Fundamentos pedagógicos de la evaluación

Guía pr áctica para educadores

José Manuel García Ramos

EDITORIAL
SINTESIS

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y el resarcimiento civil previstos en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente, por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o por cualquier otro, sin la autorización previa por escrito de Editorial Síntesis, S. A.

© José Manuel García Ramos

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso, 34. 28015 Madrid
Teléfono 91 593 20 98
http://www.sintesis.com

ISBN: 978-84-995871-9-6

Impreso en España - Printed in Spain

Índice

Introducción

PRIMERA PARTE

LA EVALUACIÓN EDUCATIVA: CONCEPTO Y PROBLEMÁTICA

1. Investigación y evaluación. Implicaciones y efectos

1.1. Los fines de la investigación científica

1.2. Ciencia

1.3. Algunos principios de la filosofía de la ciencia en la actualidad

1.4. Dos conceptos: causalidad y empirismo

1.5. El método y la cuestión de los paradigmas en la investigación y evaluación educativas

1.6. La medida en investigación y evaluación educativas

1.7. La evaluación educativa y sus ámbitos

1.8. Investigación e innovación educativa

1.9. Investigación y construcción de teoría en educación

1.10. Investigación sobre evaluación. Algunas conclusiones

2. ¿Qué es la evaluación? Necesidad de evaluar

2.1. La evaluación y su evolución conceptual

2.2. Evolución del concepto de evaluación

2.3. Algunas conclusiones sobre la evaluación y su incidencia en la calidad educativa

3. El rendimiento educativo. Acotaciones

3.1. ¿Qué es el rendimiento?

3.2. Evaluación del rendimiento: formas e instrumentos

4. Modalidades de la evaluación

4.1. Evaluaciones inicial, continua y final (momentos de la evaluación)

4.1.1. Evaluación inicial

4.1.2. Evaluación continua

4.1.3. Evaluación final

4.2. Evaluaciones formativa y sumativa. Funciones de la evaluación

4.3. Evaluaciones normativa, criterial y personalizante (sistemas de referencia en la evaluación)

4.4. Otras modalidades de evaluación

4.5. Otras funciones de la evaluación

5. La medida en la evaluación educativa

SEGUNDA PARTE

TÉCNICAS E INSTRUMENTOS DE RECOGIDA DE INFORMACIÓN Y MEDIDA

6. Instrumentos de registro de datos

6.1. Registro de datos de observación

6.1.1. Observación directa

6.1.2. Observación indirecta

6.2. Instrumentos de encuesta

6.2.1. El cuestionario

6.2.2. La entrevista

6.2.3. Técnicas sociométricas

7. Instrumentos de medición: tests y pruebas

7.1. Instrumentos externos de medida (tipificados)

7.1.1. Tests tipificados de medición de aptitudes

7.1.2. Tests de personalidad, intereses, adaptación, actitudes, estilos cognitivos

7.1.3. Tests e instrumentos externos de medición del rendimiento y de otras áreas pedagógicas

7.1.4. Valoración e interpretación de los tests tipificados

7.2. Instrumentos internos de medida del rendimiento

7.2.1. Referencia criterial y normativa

7.2.2. La prueba objetiva

7.2.3. Problemas

7.2.4. La prueba de ensayo

8. La práctica evaluadora en el centro escolar

8.1. Planificación general de la evaluación en un centro escolar

8.2. Pasos de un proceso tipo de evaluación

8.2.1. Fase de planificación

8.2.2. Fase de construcción y selección de instrumentos de recogida de información

8.2.3. Fase de recogida de datos

8.2.4. Evaluación (formulación de juicios, toma de decisiones e informes)

TERCERA PARTE

LA MEDIDA EN EDUCACIÓN. TEORÍAS DE LA MEDIDA Y PROBLEMAS RESUELTOS

9. Teorías de la medida

9.1. Teoría clásica de la medida

9.1.1. La Teoría de la puntuación verdadera

9.1.2. Algunas limitaciones de la Teoría clásica de la medida

9.2. Teoría de respuesta al ítem

10. Resolución de problemas sobre medición en educación

10.1. Teoría Clásica de la medida

10.1.1. Formulario

10.1.2. Ejercicios resueltos

10.1.3. Validez

10.1.4. Análisis de ítems

10.1.5. Tipificación

10.2. Teoría de respuesta al ítem

10.3. Teorías de medida: problemas y soluciones

10.3.1. Teoría Clásica: problemas y soluciones

10.3.2. Teoría de respuesta al ítem: problemas y soluciones

a

Introducción

Esta obra ha sido diseñada para ayudar a futuros pedagogos y a profesores en ejercicio a comprender mejor los principios fundamentales en que se basa la evaluación educativa, así como a facilitar el diseño de la propia evaluación, como herramienta y palanca del proceso de enseñanza-aprendizaje en el contexto de aplicación a una institución educativa.

El libro consta de tres partes. En su primera parte (capítulos 1 a 5), el libro pretende ofrecer las bases conceptuales en las que se apoya la evaluación educativa, con un especial énfasis en la acotación de los conceptos de evaluación y rendimiento educativos. En esta parte también se analizan las distintas modalidades de evaluación, que pueden definirse según distintos criterios de clasificación: inicial, continua y final, en clara referencia a los momentos del proceso global de evaluación; formativa y sumativa, derivadas de los objetivos que persigue la evaluación; normativa, criterial y personalizante, diferenciación establecida en función de la instancia de referencia que se utilice para evaluar (heterorreferencia o autorreferencia) y otras modalidades derivadas de otros criterios clasificatorios. También se ofrecen breves reflexiones sobre las funciones que puede cumplir la evaluación (de control, predictiva, orientadora, etc.), funciones que sin duda condicionan el modelo evaluativo elegido. Finalmente, en esta primera parte se analiza el papel de la medida en la evaluación educativa, papel central como herramienta básica para la planificación, desarrollo e interpretación de la propia evaluación. Medir no es evaluar, pero si partimos de medidas para la evaluación, tendremos que asegurar que dichas medidas aporten información fiable, consistente y no deformada por defectos en los instrumentos empleados para obtenerla.

La segunda parte de la obra (capítulos 6, 7 y 8) está dedicada fundamentalmente a la clasificación y diseño de técnicas e instrumentos de recogida de información y medida; tiene como objetivo conocer, valorar y ayudar a diseñar las herramientas metodológicas que servirán de base para la recogida de información, como paso central en el proceso evaluativo. Hemos clasificado dichas técnicas en instrumentos de registro de datos (observación directa o indirecta y de encuesta: cuestionarios, entrevistas y técnicas sociométricas), e instrumentos de medición en sentido estricto. A estos últimos, a su vez, los hemos dividido en externos (tipificados) e internos (construidos por el profesor o equipo docente) en clara alusión a la función del profesor de seleccionar (tests externos de entre los disponibles en el mercado) o de construir (pruebas ad hoc para medir variables de rendimiento). Hemos puesto especial énfasis en este último tipo de instrumentos - los internos - dado que en la mayoría de las ocasiones el profesor deberá elaborar sus propias pruebas de evaluación, procurando definir las características de cada una de ellas, así como los pasos que deberán seguirse para construir pruebas objetivas, pruebas de problemas y pruebas de ensayo. En todas ellas el sistema de referencia puede

definirse como aproximación a la norma o al criterio. El capítulo 8 ofrece algunas pautas sobre cómo realizar el proceso evaluativo: un esquema en cuatro fases y diez pasos que pretenden secuenciar y ordenar los distintos elementos integrantes del proceso evaluativo.

La tercera y última parte del libro (capítulos 9 y 10) pretende ofrecer las bases de la medida en educación. El capítulo 9 pretende ser una síntesis de las dos grandes teorías de la medida: la Teoría Clásica (TCT) y la Teoría de Respuesta al ítem (TRI), dos teorías complementarias, pero que abordan conceptos diferenciados para fundamentar una medida científica y sistemática. Por último, el capítulo 10 ofrece pautas para abordar la resolución de problemas sencillos vinculados a ambas teorías de la medida.

A Mari Carmen y a mis hijos José Manuel, Pablo, Dani, Carlos, M.a José y M.aCarmen.

PRIMERA PARTE

La evaluación educativa: concepto y problemática

1

Investigación y evaluación. Implicaciones y efectos

Es preciso reconocer que asistimos a unos momentos en los cuales se están multiplicando los puntos de vista sobre la metodología de la investigación y evaluación educativa.

Son muchos los temas metodológicos sobre los que hay y se están formando opiniones con repercusiones teóricas y prácticas.

Por ello se hace preciso situar la investigación educativa (y por ende, la evaluación) en relación con la ciencia, la tecnología, la pedagogía, las distintas disciplinas pedagógicas y la misma práctica educativa, desde un nivel conceptual y epistemológico, para pasar después a analizar las consecuencias de tales relaciones.

Estas reflexiones conceptuales nos ayudarán también a definir con mayor precisión nuestros conceptos de investigación y evaluación, así como su ámbito de aplicación en el campo educativo.

i.i. Los fines de la investigación científica

Los clásicos niveles de investigación educativa reflejan de alguna manera los principales fines: básica, aplicada y activa o experimentación.

- La primera vía es sustantiva, vital, su ausencia deja sin fundamento la pretensión de cientifidad de cualquier rama del saber, sin ella se hace dificil pensar en la existencia de los otros niveles.

- La investigación aplicada y la investigación activa se dirigen fundamentalmente a la solución de problemas o a la explicación de fenómenos y situaciones educativas.

- Cuanto más próximos a la acción concreta y contextualizada, más lejanos y ajenos a la búsqueda de teorías generales.

- Todo ello nos lleva directamente al debate epistemológico sobre la ciencia y las ciencias.

1.2. Ciencia

El término ciencia es hoy considerado un término equívoco, ambiguo, dificil de caracterizar, independientemente de cuál sea su objeto.

Las razones de ello son varias. Sin detenernos a analizarlas, en la actualidad suelen definirse como rasgos más característicos de la ciencia los siguientes:

a)La objetividad. Se trata de una característica del método científico que se consigue a través de la contrastación intersubjetiva.

b)La actitud crítica. Si algo caracteriza al científico contemporáneo es su actitud crítica. Esta actitud falsacionista es central en los planteamientos metodológicos.

c)La facticidad. Son muchos los autores que sostienen que la base empírica y la experiencia son los aspectos que mejor distinguen el conocimiento científico del no científico.

d)La racionalidad. Esta característica nos permite sistematizar de forma coherente enunciados contrastables y fundamentar una teoría sobre la realidad que la sustenta.

e)La contrastación. Contrastar una teoría es someterla a prueba de la experiencia. Se trata de un proceso que supone conjugar lo racional y lo fáctico.

f)La autonomía. La actividad científica posee un carácter primordialmente interno aunque ningún investigador carece de elementos "a priori"; es decir, todos llevamos implícita una cosmovisión subyacente que orienta de alguna manera nuestras posiciones teóricas o conceptuales. En este sentido la actividad científica es interna, pero no está exenta de planteamientos previos procedentes de esta cosmovisión.

g)Lenguaje claro y preciso. Esta característica posibilita la comunicabilidad de los contenidos, facilitando además el carácter autocorrectivo del proceso científico.

h)La sistematización. Tanto en el procedimiento como en la organización de los contenidos y conocimientos.

i)La progresión. Capacidad que la ciencia tiene para resolver problemas. Como podemos apreciar, el concepto de ciencia ha ido evolucionando hasta llegar a entenderse como una realidad compleja cuyos componentes más destacables serían:

1. Tratarse de un tipo de conocimiento más riguroso que el ordinario. Ser una actividad que comporta un método propio.

2. Poseer un lenguaje específico.

3. Aparecer como una realidad dinámica de carácter autocorrector que busca

incrementar los niveles de verosimilitud.

1.3. Algunos principios de la filosofía de la ciencia en la actualidad

Según los especialistas sobre el tema los principios más importantes que caracterizan a la epistemología actual son los siguientes:

1. Rechazo del inductivismo.

2. El problema de la verosimilitud de las teorías. No es posible confirmar las teorías.

3. Determinación de las teorías por la lógica (por el método). Nunca se puede separar totalmente conocimiento obtenido de método utilizado.

4. Condicionantes sociales de la ciencia. Importancia de las características psicológicas y sociológicas de las comunidades científicas en el progreso de la ciencia, incluso por encima de los componentes puramente lógicos de dicho proceso.

5. Relativismo en la demarcación de la ciencia. La tradicional discusión entre racionalistas y relativistas, en torno a qué es ciencia y qué no lo es, parece superada tras la afirmación de Kuhn, según la cual "no existe ninguna norma superior a la aprobación de la comunidad científica".

6. Incomparabilidad de las teorías. Si la teoría y las observaciones o hechos son inseparables, parece claro que los resultados y consecuencias de cada investigación no son separables de su contexto teórico.

Sobre la base de estos principios se establecen unas condiciones para el uso de cualquier aproximación o para la valoración de las aproximaciones paradigmáticas que actualmente se postulan.

La filosofía de la ciencia actual parece no admitir que existan criterios absolutos de demarcación del concepto de ciencia, y tampoco la existencia de conocimientos científicos como conocimientos seguros adquiridos gracias a la observación y a la experimentación y, en consecuencia, que exista un procedimiento inferencial que nos permita derivar teorías científicas de manera absolutamente fiable.

En este sentido, se afirma que todos los conocimientos deberán ser juzgados por sus propios méritos, por las soluciones que aportan dentro de un contexto social (y de una comunidad científica).

La epistemología actual parece orientarnos a aceptar abiertamente la legitimidad de las distintas metodologías de investigación educativa, evitando todo género de exclusivismo o complejos de "superioridad" o "inferioridad".

En todo caso y como investigadores, cualquiera que sea nuestra metodología prevalente o preferente, hacemos lo correcto al exigirnos unos a otros, mutuamente, rigor y calidad en los planteamientos, desarrollo y conclusiones de nuestras investigaciones.

1.4. Dos conceptos: causalidad y empirismo

El concepto de causalidad, muy consolidado en la filosofía tradicional y en muchos enfoques actuales de la investigación científica, parte de la idea según la cual "nada ocurre sin causa", todo efecto es consecuencia de una causa inmediata, pudiéndose así ordenar una sucesión de efectos y causas.

Es evidente que el pensamiento causal ha jugado y juega un papel importante en la investigación científica.

Sin embargo y especialmente en este siglo, se ha sometido a crítica tal concepto, aplicado a la realidad; incluso desde la propia física, donde podemos ver la evolución, desde la certeza absoluta de la mecánica de Laplace hasta las actuales concepciones evolutivas (catástrofe-caos) de un Prigogine, pasando por las aproximaciones probabilísticas (principio de indeterminación de Heisenberg).

Por otra parte, las restricciones y condiciones exigidas para el cumplimiento de la causalidad entre variables son tan fuertes que la misma posibilidad de establecerse parece dificil.

Sin embargo y pese a todo ello, la causalidad está implícita, cuando no explícita, en la mayoría de los postulados.

Con la aparición de los modelos causales y de los modelos de análisis de estructuras de covarianza, el concepto de causalidad es utilizado en la vía no experimental.

En esta vía el asunto es mucho más complejo y delicado, pudiéndose establecer "causalidad" por analogía, en diseños o estudios donde ni hay manipulación de variables ni azar posible, a través de la teoría.

Finalmente y para evitar la "crítica fácil" al concepto de causalidad, conviene distinguir entre dos conceptos de causalidad de diferente calado: un concepto ingenuo y simplista y un concepto más sólido y complejo.

En este sentido, conviene superar el primero, en virtud del cual una sola causa induce un solo efecto mediante un proceso susceptible de ser teóricamente representado mediante una estructura funcional sencilla y manejable, percatándose de que en muchos casos -la mayoría en nuestras ciencias sociales y en nuestra ciencia pedagógica - lo que se da es un "enjambre" o conjunto de con-causas generadoras de "enjambres" de efectos, ligados entre sí, mediante procesos que no admiten una representación teórica en

términos de estructuras funcionales sencillas.

Una cosa será admitir este segundo concepto de causalidad y no el primero, por ingenuo y otra bien distinta concluir por el primero que la causalidad, en cuanto posibilidad de intelección racional de los fenómenos, no existe.

Evidentemente, defendemos la necesidad de usar el concepto de causalidad, correctamente, pero sin complejos de ningún tipo y con toda la extensión y fuerza con que sea posible.

El término "empirismo", como el ya citado de "ciencia" es hoy un término polisémico, con muchos matices y que requiere muchas cualificaciones.

En nuestro campo y desde nuestro quehacer de investigadores, partimos de la necesidad de derivar y contrastar nuestros conocimientos con la realidad.

El dato empírico es importante, sea como base del contraste de una afirmación o para la búsqueda de acuerdo de interpretaciones.

Ciertamente, se excluyen quienes sólo hacen interpretaciones en sus aproximaciones a la realidad, sin aportar datos, ya que eliminan la posibilidad de una contrastación externa, la posibilidad de una comparación intersubjetiva, una condición ineludible en el contexto de nuestra disciplina.

r.5. El método y la cuestión de los paradigmas en la investigación y evaluación educativas

De lo anterior se deriva que, si bien la objetividad y la neutralidad no deben exigirse, por inalcanzables, sí debe pedirse en toda investigación rigor y sistematicidad, que en la mayo ría de los casos vienen dados por la preocupación, por la validez interna o lógica de nuestras investigaciones, preocupación que debe plasmarse siempre en la fase de evaluación de la consistencia interna de nuestras teorías y también por el respeto al método en la evaluación de la consistencia interna entre la teoría y los hechos.

La discusión, polémica o debate, especialmente candente en los últimos años, en torno a los paradigmas y metodologías, tiene mucho que ver con el problema del método, aunque parte quizás de un nivel de abstracción superior y estrechamente vinculado a cuestiones epistemológicas, de filosofía de la ciencia.

Las raíces del debate son antiguas. En la actualidad, hay que señalar que este debate ha generado mucha confusión. No vamos a reproducirlo aquí. Ha sido, y para algunos todavía es, un "debate candente".

Por otro lado, la controversia va más allá todavía, pues afecta a quienes piensan que las distintas aproximaciones son compatibles, frente a los que opinan que no es posible la

integración.

El problema es en parte diferente al hablar de investigación y al hablar de evaluación de programas-investigación evaluativa, en donde los objetivos son distintos, no buscándose tanto construir teoría, conocimiento más o menos generalizable, sino más bien obtener información para valorar y tomar decisiones (de mejora, de cambio, de transformación).

La complejidad del debate y del problema y las situaciones de los investigadores ante él hacen difícil enfocarlo en unas pocas páginas. No obstante, conviene distinguir brevemente dos niveles en esta discusión: planteamientos epistemológicos y metodológicos-técnicos (métodos y diseños).

El primer nivel trata de recoger las distintas posiciones, enfoques o paradigmas que dirigen la investigación y la evaluación.

Claro que, bajo las distintas definiciones de investigación y de evaluación, por lo general subyacen presupuestos filosóficos distintos acerca de epistemología y valor. Dichos presupuestos orientan y definen la naturaleza de la evaluación y de la investigación que se realiza.

Dependiendo de ellos, de cada posición, se utilizarán unos métodos de investigación y unos modelos de evaluación, así como distintos diseños/planteamientos de investigaciónevaluación, como herramientas preferentes de trabajo.

La literatura reciente está llena de argumentos acerca de la naturaleza del conocimiento. No cabe duda de que la investigación y la evaluación, especialmente esta última, se ven fuertemente afectadas por estas posiciones. De hecho, lo más o menos aceptable de una investigación concreta o de una investigación evaluativa particular depende a menudo, de la óptica que toma el que la valora.

Básicamente y en un primer intento clasificatorio, pueden distinguirse tres posiciones generales:

a)naturalista-cualitativa;

b)experimental-cuantitativa-racional y

c)ecléctica.

No nos detenemos a explicitarlos en este momento.

El criterio de verdad es básico para distinguir entre las posiciones extremas del continuo (en este nivel).

Un segundo problema es si: ¿Pueden integrarse? ¿Son complementarios? ¿Son compatibles?

Hay tesis para todos los gustos, desde las tesis de la oposición o incompatibilidad epistemológica, hasta la tesis de la complementariedad, pasando por la tesis de la unidad epistemológica con diversidad metódica aceptada. Muchos otros autores han defendido la complementariedad, pero desde posiciones y planteamientos muy distintos y dispares.

Mi posición personal en este punto es la de procurar evitar el caer en la renuncia a la construcción de una teoría sistemática y ello sobre la siguiente tesis: en un proceso racionalobjetivo no tiene por qué rechazarse el que muchas variables que han de ser integradas en la teoría no sean ni cuantificables, ni siquiera fácilmente representables mediante categorías analíticas sencillas; pero esto no tiene por qué conducimos a posiciones en las que parece renunciarse a establecer una teoría sistemática; por el contrario, a lo que debe conducirnos es a un amplio esfuerzo de producción conceptual y analítica, progresivamente conducente al establecimiento de teorías más ricas y más complejas.

En este sentido, no conviene olvidar que una de las mayores acusaciones que se hacen a la investigación educativa es la de que no se apoya en la teoría.

En el segundo nivel, las consecuencias de los planteamientos epistemológicos son evidentes; conviene, en todo caso, no confundir el debate paradigmático con el debate metodológico cuantitativo-cualitativo; aunque es verdad, que para entender cómo se llegó al segundo debate haya que tener en cuenta el primero.

Así, por ejemplo, positivismo y cuantitativismo no son la misma cosa, no conviene identificar paradigma con metodología.

En este nivel, mantenemos nuestra posición anterior, a la que conviene añadir un matiz importante: probablemente, la fecundación progresiva de la teoría dentro del marco en que nos movemos, esto es, la búsqueda de una teoría compleja progresiva exige la aportación de elementos que hoy quedan objetivamente fuera del horizonte conceptual perceptivo de las diferentes posiciones teóricas o argumentales propias de nuestro ámbito educativo, elementos estos que muy probablemente deban de ser tomados de una amplia reflexión, por ejemplo, antropológica.

El tema de los métodos en el ámbito de la evaluación educativa presenta algunas variaciones importantes respecto al ámbito de la investigación educativa.

Aun cuando la metodología de la evaluación de programas proviene claramente de la metodología de investigación, ha tenido que ir haciendo adaptaciones sucesivas para resolver problemas específicos que han ido dando lugar a métodos propios que, siendo

básicamente los mismos en su proceso a los de la investigación aplicada, introducen algunas variantes o particularidades.

En la actualidad, la evaluación de programas se caracteriza por la falta de acuerdo en lo que es, en cómo debe hacerse y para qué debe hacerse la propia evaluación.

De hecho, podríamos volver a agrupar los grandes enfoques metodológicos alternativos, de nuevo, en dos principales:

1. Esquema tradicional-objetivista y básicamente externo y

2. Modelo participativo-subjetivista-crítico y básicamente interno;

Estas dos alternativas nos devuelven a la discusión originaria que ya hemos comentado.

En este ámbito de la evaluación de programas, investigación-evaluativa, se recogen, dentro de sus diseños, técnicas y herramientas al uso, todas las posibilidades que ofrece la investigación clásica en ciencias sociales más algunas otras posibilidades, más o menos novedosas y las procedentes de la corriente cualitativa, que han ido surgiendo en la corta historia, todavía, de este área de la investigación, o si se prefiere, en este área evaluativa.

En todo caso, no debe olvidarse que una característica de la investigación evaluativa, que la distingue de la mayoría de las demás clases de investigación, es la de que se realiza en un marco de acción.

La investigación es aquí un apéndice, un algo que "también está presente", algo que tiene sin embargo, prioridad secundaria.

Con frecuencia, los investigadores proponemos cambiar el orden de prioridades y con algo de razón, pero ello puede ser un grave error.

En todo caso, el investigador evaluativo, el evaluador de programas, debe conocer mucho sobre investigación, conocer lo que se encuentra en los textos acerca de métodos de investigación y luego tiene que aprender a aplicar tal conocimiento a un ámbito que a menudo resulta inhóspito para áreas importantes de su conocimiento.

> "Si luego no sale de su libro de texto, corre el peligro de hacer un trabajo que nada tenga que ver con lo que se le encargó, de "chocar" con el personal del programa o del centro, con el que trabaja y de ver desperdiciados los resultados del estudio, si es que llega a completar su trabajo" (Weis, 1975).

1.6. La medida en investigación y evaluación educativas

Un aspecto esencial que conviene destacar es el de los instrumentos de medida, y más importante aún el de la medida misma, como base y elemento esencial e inseparable de los procesos de investigación y evaluación educativos. A este tema dedicamos gran parte del libro.

Medir es justamente el procedimiento para definir, obtener y ofrecer datos, información útil que posibilite el análisis para la obtención de conclusiones en todo proceso de investigación, así como el juicio o valoración posterior en que toda evaluación consiste.

También en el ámbito de la medida ha tenido un fuerte eco el debate cuantitativo-cualitativo.

Sin volver al debate, hay que reconocer que la preocupación por los datos y su calidad, en estos últimos años y en ambas corrientes metodológicas, es grande.

Mientras en la perspectiva cuantitativa se siguen dos direcciones:

a)mejoras en las teorías y métodos de aplicación; y

b)profundización en las teorías formales.

En la perspectiva cualitativa se busca suplir la imperfección de las aproximaciones a la realidad e ir aumentando su validez, corrigiendo fuentes de error, por la reflexión y análisis cuidadoso de las condiciones.

En mi opinión, el concepto de medida, aceptando las clásicas definiciones de distintos autores... debe ligarse a un concepto más amplio que el puramente operacional debiendo vincularse siempre a una teoría sustantiva o sistema de proposiciones referidas a un conjunto de constructos, que sirven para describir, explicar y predecir algunos aspectos limitados de su dominio.

La definición de Lord y Novick (1968) es una magnífica síntesis de lo anterior:

> "La medida es un procedimiento para la asignación de números - puntuaciones, medidas - a propiedades especificadas de unidades experimentales, de modo que las caractericen y preserven las relaciones especificadas en el dominio comportamental".

De ahí que la medida comience siempre con un procedimiento para identificar y relacionar los elementos del mundo real con elementos o constructos del sistema lógico, por medio de una definición semántica precisa; una vez hecho esto, puede procederse a la medición.

La medición educativa se enfrenta con grandes problemas teóricos y prácticos,

derivados todos ellos de las dificultades de construir instrumentos adecuados a los constructos que se definen, muchos de los cuales presentan graves dificultades de delimitación.

De aquí que se siga diciendo que el principal problema que ha de enfrentar un instrumento de medida es el de la validez de constructo.

1.7. La evaluación educativa y sus ámbitos

Con relación al concepto de evaluación y sus ámbitos, en primer lugar conviene señalar la evolución que ha sufrido el mismo concepto de evaluación a lo largo de los últimos setenta años.

Ha evolucionado rápida y profundamente, desde la conceptualización de Tyler hasta nuestros días. Inicialmente estaba restringido - en el ámbito educativo - por la referencia al aprendizaje-rendimiento de los alumnos. Actualmente y como consecuencia de su desarrollo histórico, el concepto de evaluación se ha ampliado enormemente.

Toda actividad educativa intencional es susceptible de evaluación y por tanto es evaluación educativa.

De hecho, prácticamente todas las definiciones que sobre el término se han dado pueden englobarse en alguna de las siguientes:

- Definiciones como las de Bloom (1975), Gronlund (1985), etc., van en la línea de la evaluación de productos educativos individuales (Assessment).

- Definiciones como las de Phi Delta Kappa (1971), Stufflebean y otros (1971 y 1984), Berk (1981) van en la línea de la evaluación de programas (Evaluación).

Otras definiciones tratan de abarcar ambos conceptos (Pophan, 1980; De la Orden, 1982; Pérez Juste, 1986); incluso se puede apreciar dicha orientación en Scriven, 1967 y otros muchos autores.

La tendencia actual es considerar por separado las dos facetas del término evaluación y ello, fundamentalmente, porque implican procesos y metodologías bastante distintos.

En este sentido, la diferencia básica entre ambos conceptos parece estar en la unidadobjetivo de análisis: un individuo o sujeto único ("assessment") o bien, un programa o modalidad de intervención ("evaluation").

i.8. Investigación e innovación educativa

Respecto a la relación entre investigación e innovación educativa lo primero que conviene

apuntar es que ambos conceptos están ligados, se apoyan y justifican entre sí.

La investigación educativa, en último término, se legitima por su compromiso con la praxis educativa, por el compromiso con el perfeccionamiento de la acción educativa y, por tanto, por la repercusión que tiene en la optimización del aprendizaje a través de la innovación.

Por otro lado, conviene que toda innovación se apoye en algún tipo de resultados de investigación; por lo menos, toda innovación que pretende permanecer y consolidarse como modo sistemático de intervención educativa perfectiva.

Sin embargo, de todos es conocida la poca repercusión que, hasta el presente, ha tenido la investigación en la práctica educativa.

Este tema es y ha sido objeto de abundante literatura en los últimos años.

No vamos a reproducir aquí, de nuevo, la discusión sobre las causas que han provocado esta situación y sus consecuencias sobre ambos conceptos y sobre el debate cuantitativocualitativo.

Lo que sí parece claro es que se buscan formas de superar esta situación y sobre todo se observa una creciente preocupación por conectar adecuadamente investigación e innovación, investigación y cambio educativo, innovación y evaluación, etc.

Sin duda, la vía metodológica es esencial, pues a través de la mejora y adecuación de los métodos de investigación podremos conectar adecuadamente investigación e innovación.

1.9. Investigación y construcción de teoría en educación

Una de las cuestiones que surge siempre en nuestro ámbito educativo es la consideración de si la Pedagogía es ciencia o no lo es.

Ciertamente, el quehacer pedagógico-educativo es un arte, la cuestión es si además es una ciencia.

Tal como definimos la ciencia anteriormente y de forma amplia, la Pedagogía es también una ciencia, en cuanto que utiliza métodos que se caracterizan por tener valor heurístico y ser acción reflexiva, rigurosa, sistemática, organizada y disciplinada.

La Pedagogía tiene modos de conocimiento rigurosos y metódicos con el fin de descubrir leyes en el objeto de estudio y de expresar los conocimientos adquiridos sistemáticamente. Es decir, tiene objeto, rigor, método, leyes y teorías sistemáticas. Ciertamente, lo que más cuesta aceptar en Pedagogía es que tenga leyes; pero el

problema se reduce a la cuestión de a qué nivel de generalidad-universalidad es necesario que se formulen enunciados de la Pedagogía para que puedan considerarse leyes científicas.

Citando a Sherwood (1969), quien presenta los distintos tipos de generalización aceptables en la ciencia, puede afirmarse que la Pedagogía está consiguiendo leyes y teorías progresivamente más generales-universales.

Por otro lado y en cuanto que los conocimientos pedagógicos se pretende tengan un alto grado de practicidad, al permitir a más corto o más largo plazo, su aplicación a la mejora educativa, puede ser considerada una ciencia tecnológica.

Se trata pues, de la mejora del sistema educativo, de la educación en general, como criterio de validez (cuestión ésta no exenta de valores de enjuiciamiento).

Lo que habrá que evitar es que al enfatizar el carácter práctico del saber pedagógico, descuidemos y olvidemos que los fenómenos educativos, objeto de estudio de la Pedagogía, por complejos y multivariados que sean, son susceptibles de conocimiento sistemático integrable en teorías.

Los métodos de investigación y evaluación en Educación se caracterizan por las siguientes notas:

a)Finalidad científica.

b)Rigor en sus procedimientos.

c)Énfasis en el método.

d)Carácter teórico-práctico.

e)Perspectiva o enfoque de los problemas educativos.

f)Metodología empírico-experimental; metodologías de base empírica o experimental que buscan el contraste científico de la realidad.

Ahora bien, aceptando todo lo anterior, no es menos cierto que el gran drama que las ciencias sociales (Economía, Psicología, etc.) tienen planteado en este momento y la Pedagogía no es una excepción (en las Ciencias Naturales y en la Física por ejemplo, también ocurre aunque en menor medida) es el haber descubierto que la mayoría de las variables explicativas de los fenómenos con que trabajamos son variables "cualitativas" (entendiendo por ellas las que comúnmente denominamos nominales y especialmente, ordinales) y no tanto variables "cuantitativas" (susceptibles de ser medidas cardinalmente, lo que comúnmente denominamos como de intervalo o de razón).

Ello nos obliga a definir teorías explicativas cada vez más generales, más complejas, más potentes en el sentido popperiano, en las cuales estén presentes ambos tipos de variables.

Las primeras, no cuantificables estrictamente, son la mayoría en nuestro ámbito educativo.

Aun aceptando la medida de variables como rendimiento, aptitudes y algunas otras, al nivel de cuasi intervalo "cuantificables", la mayoría siguen siendo ordinales o nominales.

Estas variables son, sin embargo, susceptibles de algún tipo de medida, por supuesto son susceptibles de ser definidas en términos lógicos y, sobre todo, presentan alto poder lógico-explicativo de muchos fenómenos educativos.

En este sentido, entiendo que nuestra disciplina debe conducirnos a trabajar con Teorías Tipo 2 (T2), según el esquema siguiente:

-TI: Sólo maneja categorías analíticas cuantificables (k), por tanto, hipótesis con sólo variables k, que a través del proceso inferencial desembocan en Teoremas K, que son susceptibles de contraste empírico.

-T2: maneja todo tipo de categorías analíticas que puedan ser acuñadas para caracterizar razonablemente un grupo de fenómenos; algunas serían k y otras no, serían q (cualitativas).

Por tanto, tendríamos hipótesis Ke hipótesis Q, orgánicamente ligadas, que a través del proceso inferencial desembocan en teoremas que implican relaciones "funcionales" o estructurales que son contrastables empíricamente en términos de k y con implicaciones k, sobre condiciones definidas por las categorías q.

Ciertamente, en todo caso habremos de definir y justificar razonablemente, qué variables consideramos k y cuáles q.

Por otro lado, y asumiendo la ampliación que en los últimos años se ha producido en la disciplina que con la denominación de "Métodos de investigación" en Educación ha incorporado otros métodos de investigación además del enfoque tradicional y el concepto de ciencia que hemos defendido anteriormente, parece lógico afirmar que nuestra disciplina en la actualidad se caracterice por su matiz metodológico, que abarcaría así las distintas aproximaciones empírico-experimentales, cuyos objetivos son: la descripción, explicación, predicción y control, de los fenómenos educativos.

Una definición que podría ayudamos a resumir todo lo dicho hasta aquí nos llevaría a decir que entendemos la investigación en educación, como el estudio de los métodos y metodologías, técnicas y procedimientos utilizados y utilizables, para obtener un

conocimiento y una explicación científica de los fenómenos educativos, así como para mejorar la realidad educativa, mediante una aproximación técnica y rigurosa a los problemas educativos y sociales, que tal realidad plantea.

Por otra parte y sobre la base de todo lo dicho hasta aquí, me gustaría concluir con algunas propuestas personales sobre cómo superar el actual debate metodológico.

Para ello, se hace necesaria la propuesta de una metodología de investigación razonable, que yo apoyaría en mi propuesta anterior de trabajar con Teorías Tipo 2. Ello significa que:

a)Toda teoría debe proponer relaciones causales, sea dicha teoría cualitativa o cuantitativa. En este sentido, toda teoría es, por definición, hipotético-deductiva.

Conviene recordar aquí que la inducción no puede aceptarse, pues los enunciados observacionales están cargados de teoría, destruyéndose así la tesis de que la ciencia comienza con la observación.

En esta línea, la llamada "vía inductiva" que prefiere que la "teoría emerja de los datos" no parece aceptable.

Como sabemos, los datos no existen independientemente de la estructura teórica de alguien.

Algo es erigido en dato por alguien que lo hace a partir de alguna clase de estructura teórica (buena o mala, consciente o inconsciente; explícita o implícita) que le permite reconocer como dato algo que percibe.

Como ejemplo, basta decir que yo no puedo distinguir entre un "esquizofrénico" y un paciente con "delirio paranoico". Sin embargo, un buen psiquiatra sí que puede. De la misma manera que no soy capaz de distinguir si se ha estropeado el cárter o una bujía, por el ruido del motor del coche. Un buen mecánico puede hacerlo.

Y es que para mí, los síntomas del enfermo o el ruido del motor no son datos, son hechos o fenómenos que no reconozco como datos mientras no tenga una estructura teórica mínima que lo permita. No puedo o no soy capaz de discernir o de aceptar como significativo un dato que no puedo identificar como tal.

Ahora bien, esto no quiere decir que la percepción desde una teoría inicial (TO) no pueda plantearme un campo de extensión conceptual o tensión problemática con respecto a TO.

En ese caso, no es la mera percepción de algo distinto - sin más - lo que hace que esa tensión problemática desaparezca, sino la reelaboración hipotético-deductiva a

partir de TO; ésta es la que me conduce a una reflexión a partir de un nuevo dato.

Partimos así de la existencia siempre de una teoría fundamental, entendida como conjunto, frecuentemente no bien articulado, de principios epistemológicos y sustantivos (teóricos) que tenemos por ciertos y a través de los cuales tematizamos nuestra observación.

b)Ciertamente, la realidad es objetiva, pero no lo es la visión de la misma desde nuestra subjetividad.

La visión de la realidad está siempre mediada por la subjetividad.

De hecho, todos podemos constatar que sobre una misma realidad uno ve cosas que otros, simplemente, no ven y viceversa.

Parte sustantiva de esta propuesta personal de superación del debate metodológico es pues, el reconocimiento de que necesitamos Teorías, debiendo dirigir nuestro esfuerzo principal a proponerlas y después, contrastarlas con la realidad.

Estas teorías, como ya se dijo, manejarán variables cuantitativas (k) y cualitativas (q) que permitan ligarlas orgánicamente y que a través del proceso inferencial (hipotético-deductivo) desemboquen en el establecimiento de relaciones funcionales o estructurales que sean contrastables empíricamente.

Ciertamente, nuestra mayor dificultad es que trabajamos, en la mayoría de las ocasiones, con variables cualitativas.

Ello nos obligará a un gran esfuerzo de producción teórica, conceptual y analítica, progresivamente conducente a teorías más ricas y más complejas.

En cierto sentido, puede decirse que la mayoría de nuestras teorías son o serán cualitativas y deberán constituirse por la vía hipotético-deductiva en continua interacción con la realidad que pretendemos explicar, comprender y mejorar. Lo mismo ocurre en otras ciencias sociales, como por ejemplo la Economía, donde todas las teorías actuales trabajan con variables cualitativas y proceden del trabajo desarrollado por vía hipotético-deductiva.

No obstante, ello no implica reducir las variables con las que trabajamos a variables únicamente cuantitativas (k), tal como comentamos al hablar de Teorías tipo 1 y tipo 2.

Si lo hiciéramos, estaríamos dejando fuera del campo explicativo de la teoría fenómenos en los que intervienen variables "q" (la mayoría, como hemos dicho). No admitiríamos así una posición como ésta sistemáticamente reduccionista.

El clásico ejemplo del "farol y la moneda" puede iluminar lo que aquí defendemos.

El farol y la moneda

En este clásico ejemplo, muy usado de todas las ciencias sociales, una persona pierde una moneda al cruzar la calle.

Al llegar al otro lado se da cuenta de ello. En la calle sólo hay un farol, que ilumina una parte pequeña de la calle. La persona busca la moneda sólo en la zona iluminada por la farola y así lo hace durante horas.

Llega un amigo y le pregunta que qué hace. El "buscador" de la moneda dice que la busca desde hace horas y que no la encuentra. El amigo le invita a buscarla por toda la calle, negándose el "buscador", pues no está iluminada por el farol.

La moraleja es muy sencilla: habrá que "arrancar la farola, sin desconectarla de la fuente de electricidad, si hace falta" o buscar un "grupo electrógeno o una linterna potente" para buscar la moneda, sin reducir la búsqueda a la zona iluminada por el farol.

Habrá pues que acomodar el método al objeto de estudio que es la realidad objetiva, evitando reduccionismos como el señalado.

Nuetras teorías, progresivamente, deberán tratar de explicar toda la realidad.

Para ello, frecuentemente, deberán trabajar con todo tipo de variables: cuantitativas y cualitativas.

De ahí que el "enfoque cuantitativo" no sea aceptable si implica una reducción como la descrita.

Ahora bien, ello no significa - como en ocasiones se observa - que debamos identificar gratuitamente estructura hipotético-deductiva y estrategia reduccionista (ya descrita).

Los que así lo hacen creen o postulan que la superación del reduccionismo también implica el abandono de estructuras teóricas hipotético-deductivas, cayendo así en una postura difícilmente compatible con el estudio racional y sistemático de los fenómenos.

En la medida en que "investigación o enfoque cualitativo" signifique el no establecimiento de relaciones causales, racionalmente discernibles, no nos parece aceptable.

Lo que propugnamos, pues, es un enfoque progresivo en el que, asumiendo los elementos positivos de ambos enfoques, se vaya a la construcción de modelos

explicativos que combinen variables "q" y "k", que establezcan relaciones causales entre ellas, lo que permitirá una explicación (intelección y comprensión son, en este sentido, sinónimos de explicación) que naturalmente debe poder estar abierta a varios tipos de contrastación, en algunos aspectos o dimensiones cuantitativas... y en otras cualitativas.

La tarea no es fácil, ciertamente, pero el objetivo justifica el esfuerzo que entre todos debemos realizar.

i.io. Investigación sobre evaluación. Algunas conclusiones

En la actualidad pues, nos encontramos en una época caracterizada por una visión más equilibrada de las posibilidades de la investigación educativa.

Un período "de evaluación crítica", en el cual se intenta evaluar la calidad de la investigación pedagógica y su contribución al progreso de la educación.

Una época caracterizada por una gran preocupación metodológica, una época de "reflexión" sobre lo conseguido y lo no conseguido, intentando definir:

a)Cómo usar mejor los métodos y técnicas existentes.

b)Qué áreas o problemas educativos han sido y no han sido suficientemente abordados desde la metodología de investigación.

c)Nuevos modos de adaptarse al estudio de los problemas.

d)Nuevas formas de definir problemas educativos, susceptibles de ser investigados desde distintas alternativas metodológicas.

De la misma manera, podríamos afirmar que la vía metodológica y de investigación es esencial para la evaluación educativa.

En primer lugar, no debemos olvidar que Investigación y Evaluación son dos cosas distintas, que ciertamente comparten muchos elementos de proceso y método, pero con un objetivo diverso: la evaluación conduce a decisiones; por el contrario, la investigación conduce a conclusiones.

En la evaluación de programas o su correlato metodológico, la investigación evaluativa, la investigación es secundaria, lo prioritario es valorar el programa, alumno, centro, como fuente de toma de decisiones.

Con ello se quiere indicar también que la evaluación debe adaptarse al contexto y al objeto evaluado (alumnos, programas, centros, etc.) y perturbar lo menos posible las operaciones y dinámicas del mismo. Ello no significa que no se produzcan interferencias

en ocasiones.

Lógico, entre otras cosas es preciso recoger datos, pero habrá que procurar, aquí especialmente, recabar sólo los necesarios (en ocasiones, el investigador o el evaluador solicitan más información de la que utilizarán después).

En segundo lugar, debemos tener en cuenta algunos otros problemas que habrá que cuidar en la investigación evaluativa, que son:

a)La tendencia del alumno, del programa o del centro, a cambiar mientras está siendo evaluado;

b)Las relaciones entre evaluadores y el personal del programa del centro; y el hecho de que el programa se halla inserto en un sistema de organización o el centro educativo en un sistema sociopolítico más amplio y de que la naturaleza de dicho sistema condicionará, de alguna manera, los resultados.

La metodología de investigación y el sentido común pueden ser buenas herramientas para ayudar a minimizar esta problemática.

c)Dada la importancia que concedemos a la evaluación que, no lo olvidemos, tiene una influencia decisiva y directa sobre la calidad de la educación, contar con un claro sistema evaluativo es básico.

Si la evaluación se constituye en motor y palanca del aprendizaje y de la mejora de la calidad, es porque "determina en gran medida - al prescribir los objetivos de la educación - las características de la enseñanza y del aprendizaje, lo que los alumnos aprenden y cómo lo aprenden, lo que los profesores enseñan y cómo lo enseñan, los contenidos y los métodos; en otras palabras, el proceso y el producto de la educación" (De la Orden, 1982).

De aquí la importancia de cuidar la coherencia entre los objetivos de un programa o centro educativo (las metas) y las características del esquema evaluador.

No es baladí la preocupación de los responsables de los programas o centros educativos, por la evaluación.

No evaluar los programas o hacerlo intuitivamente conlleva riesgos excesivos.

Evaluar de cualquier forma es poner todos los medios para dirigir la enseñanza y el aprendizaje hacia metas desconocidas y seguramente no deseadas.

En todo caso, es necesaria la voluntad "política" de evaluar y no sólo eso, junto a ella es preciso definir: decisiones a tomar (uso de los datos evaluativos), criterios de

valoración y objetivos del programa.

Surge así uno de los problemas más peliagudos: la elección de criterios de evaluación.

En los primeros años de la todavía corta historia de la evaluación de programas, la evaluación prestó poca atención al rol de los valores.

Scriven (1983) sugiere que ello fue debido a que los evaluadores ingenuamente creían que sus actividades podían y debían ser "libres de valores".

Pero muchos evaluadores aprendieron de la experiencia, que era imposible, en el mundo político de la programación social y educativa, tomar decisiones sin tener valores que fueran sobresalientes y quizás más claramente cuando los criterios evaluativos fueran seleccionados y justificados.

Demasiados evaluadores e investigadores se han formado y entrenado en la creencia de que los valores no son parte de la ciencia.

d)Una pregunta básica que siempre debemos hacernos, al planteamos una evaluación de programas, es la de si ésta está justificada siempre.

Como afirma Weis (1975) "por herético que parezca, la respuesta es no. La evaluación, en su calidad de investigación aplicada, está consagrada al principio de utilidad". Si no tiene ninguna influencia en las decisiones, resultará una tarea vana.

La misma autora señala cuatro clases de circunstancias en las cuales no vale la pena proceder a realizar una evaluación (págs. 23-24):

1. Cuando no hay nada que preguntar acerca del programa. Se está realizando y las decisiones sobre su futuro no se van tomar o ya se han tomado.

2. Cuando el programa carece de una orientación clara. El personal del programa improvisa sus actividades de un día para otro, basándose en pocas reflexiones y menos principios y el programa cambia y se va transformando, "vagabundea" y se orienta como puede. Hay poca razón para llamar a esto "programa".

3. Cuando las personas que deberían saberlo no pueden ponerse de acuerdo en qué es lo que están tratando de realizar en el programa. Si existen grandes discrepancias en las metas pensadas, la evaluación carece de fundamento sobre el cual levantarse.

4. Cuando no hay suficiente dinero, ni personal lo bastante cualificado para resolver la evaluación. Esta es una tarea exigente, que requiere tiempo, dinero,

imaginación, tenacidad y destreza.

Aunque algunas personas pudieran argumentar que en circunstancias como las anteriores "mejor es hacer algo que nada", la experiencia demuestra que el estudio que se realiza así rara vez arroja resultados o "luces", sino que más bien acaba convirtiéndose en "papel mojado" y termina siendo arrinconado o arrojado a la papelera.

Es, pues, esencial plantearse el porqué y el para qué de la evaluación de un programa.

A veces uno puede observar que la decisión positiva sobre la realización de una evaluación particular obedece a razones poco legítimas como, por ejemplo, algunas de las señaladas por Weis (1975):

- Aplazar una decisión. En vez de enviar la "patata caliente" de la decisión sobre un programa a una comisión y esperar su informe, se encarga un estudio evaluativo que requiere aún más tiempo.

- Eludir responsabilidades. En ocasiones grupos o personas responsables de la organización del programa tienen criterios opuestos. Los gestores o administradores toman la iniciativa de pedir un estudio evaluativo que permita tomar decisiones con objetividad evitando apasionamientos. Sin embargo, en muchos casos los administradores ya tienen tomada la decisión antes de llamar a los evaluadores, pero desean revestirla del "ropaje legitimador" de la investigación-evaluación.

- Relaciones públicas. A veces la evaluación es considerada como un modo de darse "auto-bombo". El administrador cree que cuenta con un programa exitoso y busca la manera de hacer visibles sus "glorias". Un buen estudio "les vendrá de perlas". Copias del informe serán elevadas a todas las instancias superiores, consiguiendo así efectos "autoglorificantes" o económicos, si se persigue la financiación del programa.

Suchman (1970) habla de dos finalidades emparentadas: "echar humo en los ojos" y "adornarse".

En el primer caso se intenta justificar un programa débil, seleccionando para la evaluación únicamente aquellos rasgos que, superficialmente, tienen buen aspecto.

La evaluación para "adornarse" trata de encubrir el fracaso de un programa evitando toda estimación objetiva.

Por supuesto, los motivos del administrador no tienen por qué ser torcidos o egoístas. Por lo general, es necesario conseguir apoyos para un concepto o un proyecto en cuyas virtudes se cree. La evaluación puede ofrecer gran parte de esos apoyos.

•Cumplir los requisitos de una financiación externa o social. Muchas decisiones políticas de financiación de un programa exigen como requisitos la evaluación del mismo.

Desde el punto de vista de los que conceden ayuda económica, corriendo el riesgo de financiar un proyecto no probado, es razonable exigir que se aporten datos del grado en que el proyecto funciona bien.

En este sentido, la evaluación corre el riesgo de ser "un ritual necesario" para convencer o "aplacar" al financiador, pero carente de utilidad real.

Todo lo anterior nos hace pensar que la evaluación puede ser una empresa racional que a menudo se efectuará por razones que no son racionales. Ello no debe hacernos desconfiar de ella, pero debe prevenirnos de cualquier tipo de justificación. No todas son legítimas y es siempre necesario ponderar las razones que hacen que el evaluador acometa su tarea. Para empezar, será necesario preguntarse: ¿quiénes esperan obtener respuestas? y ¿qué esperan obtener?

En resumen, la investigación evaluativa, en opinión de quienes la practican, es una manera de aumentar la racionalidad de las decisiones. Téngase en cuenta que, al contar con información objetiva acerca de los resultados de los programas, es posible tomar decisiones atinadas en materia de asignación de partidas presupuestarias y planificación de los programas. Los programas que rindan buenos resultados se ampliarán, los que no los rindan serán desestimados o modificados drásticamente.

Por otra parte, sabemos que, si no se evalúan sistemática y objetivamente los programas de acción que se pongan en marcha, es fácil que los mismos programas sean desechados o ampliados por criterios "políticos" y maniobras en beneficio de grupos o personas particulares. Los datos, si se utiliza la investigación evaluativa, sustituirán a los favores y a otras negociaciones políticas, de manera que se tomarán decisiones más racionales.

No obstante lo anterior, es verdad que la breve historia de la investigación evaluativa no ha sido demasiado fructífera: es más, algunos autores la consideran decepcionante. Pocos ejemplos de aportaciones importantes a la política o los programas pueden citarse. Ello se debe, en parte, a las notables resistencias de las organizaciones a ser evaluadas, a que sus programas sean evaluados, a ofrecer la información que se precisa en la forma en que se necesita y al cambio no deseado.

Otra parte de la culpa estriba en la manera en que a veces está estructurada, dotada de personal y ejecutada la misma evaluación. En ocasiones hay incoherencia entre los fines propuestos en la evaluación y el tipo de estudios o programas evaluados.

Gran parte de la decepción que venimos comentando se debe al carácter poco realista de las expectativas que se ponen en las aportaciones que ofrecerá la investigación evaluativa. Por lo general, un estudio de evaluación no desemboca en la obtención de datos decisivos e inequívocos acerca del valor de un programa. Sus resultados muestran a menudo pequeños cambios, o ambiguos, efectos "de poca monta", resultados en los que han influido sucesos específicos de lugar y momento.

Tal vez se necesitará un estudio continuado a lo largo del tiempo y comparando proyectos, para hablar con seguridad de éxito y de fracaso.

Además, para quienes toman decisiones, la evaluación de la evidencia obtenida es sólo una "entrada" entre muchas otras.

Deben considerar otros muchos factores: receptividad del público, reacción de los participantes, costo, cantidad de recursos y de personal de que se disponga, alternativas existentes, etc.

Por otra parte, la puesta en marcha de un programa eficaz puede no aceptarse en una determinada comunidad o colectivo por considerarse que viola supuestos y valores aceptados por ese grupo.

2

¿Qué es la evaluación? Necesidad de evaluar

En el contexto educativo se vienen utilizando tres vocablos para dar a conocer y valorar, con la mayor precisión posible, el producto educativo, el producto de la actividad escolar o académica: son los conceptos de evaluación, calificación y medida.

a)El primero de ellos - evaluación - es el más general y, apoyándose en los otros dos conceptos - calificación y medida-, pretende valorar el resultado del trabajo escolar, el resultado del proceso de enseñanza-aprendizaje, la eficacia de un centro educativo, la eficacia relativa de un programa de acción; apreciando, asimismo, todas las variables que inciden en el proceso educativo: alumnos, profesores, programas, estímulos educativos, ambiente social, etc. El término evaluación, no obstante, se ha incorporado hoy en día a todos los sectores de la actividad humana y no sólo a la actividad educativa. Su utilización es, pues, universal.

La evaluación, en general, aspira a conocer y valorar, no sólo los resultados conseguidos, sino también la correlación que existe entre éstos y los medios utilizados.

La actividad evaluadora es una característica inherente a toda actividad humana intencional. Se presenta como una fase de todo sistema operativo, vinculada a los restantes elementos del mismo. De hecho, y durante décadas, la literatura pedagógica, en particular, ha considerado a la evaluación como algo necesario.

La evaluación educativa intenta conocer los resultados de las instituciones escolares, de sus programas de enseñanza, de sus métodos, etc., matizándola con criterios de valor. Como afirma Pophan (1980), "evaluar algo es determinar su valor" (pág. 16).

Así pues, y para empezar, la evaluación es un juicio de valor apoyado en conocimientos y datos de lo evaluado. La evaluación exige un conocimiento previo de unas escalas de valor que sirvan de marco de referencia a la formulación de nuestros juicios. La primera consecuencia que se deriva de ello es que evaluar exige medir, sin medidas no es posible calificar ningún resultado o producto y mucho menos evaluar en sentido objetivo, sistemático y científico.

Para muchos autores, pues, la evaluación es la valoración que se realiza sobre la base de medidas previas, medidas objetivas, precisas, fiables y válidas en cierto grado.

Ahora bien, la evaluación, siendo un elemento permanente en toda situación de aprendizaje, y en cuanto tal, implica bastante más que la simple aplicación de pruebas objetivas o algún tipo de medida. Como afirma De la Orden (1982), "el profesor deberá dirigir su atención a cuestiones como éstas: ¿logré mis objetivos?, ¿cambió la conducta de los alumnos?, ¿fueron efectivos mis procedimientos?, ¿los usaría de nuevo en una situación similar?, ¿qué destrezas, aptitudes o conocimientos necesitan más atención?, ¿qué estudiantes en mi clase necesitan más atención? En una palabra, es necesario interesarse en averiguar si se siguieron los mejores caminos para lograr los objetivos establecidos" (pág. 14).

La evaluación es pues, en sí misma, un proceso continuo y un elemento, a la vez, de toda actividad educativa intencional.

b)El segundo término, calificación, es similar al de evaluación, en cuanto que exige una comparación entre la realidad a evaluar - sujeto, objeto, situación, proceso - y el patrón o criterio de evaluación; ahora bien, menos amplio, de menor nivel de generalidad.

Por lo general, el término calificación suele utilizarse para valorar solamente la conducta de un alumno. Se habla por lo común de calificaciones escolares. Concretando un poco más, una calificación escolar es siempre una apreciación del rendimiento de un alumno en algún aspecto importante de la actividad escolar.

En la aplicación práctica de cualquier modelo evaluativo, el concepto de calificación es un elemento básico y fundamental que permite expresar cualitativa o cuantitativamente el juicio global o específico, que corresponde, con mayor o menor fiabilidad y objetividad, al resultado de la actividad del estudiante. En este juicio, se suele querer expresar el grado de suficiencia o insuficiencia de los conocimientos, destrezas o habilidades demostradas por un alumno, como resultado de algún tipo de prueba, examen o ejercicio.

La calificación escolar, pues, como apreciación del rendimiento, permite responder a algunas exigencias capitales de todo sistema evaluativo del rendimiento escolar, tales como:

-Orientar al alumno, al informarle de su proceso.

-Proporcionar al profesor un conocimiento objetivo del alumno y de la clase, a fin de que pueda adaptar la actividad a los intereses, necesidades y ritmo del

alumno.

-Facilitar el diagnóstico de los escolares, de cara a su posterior elección académica o profesional.

-Informar o motivar a los padres, ya que al indicarles el ritmo de progreso de sus hijos se les impulsa a colaborar eficazmente en las tareas formativas realizadas en la escuela.

-Suministrar datos que permitan ayudar a la planificación de la enseñanza.

c)El tercer término aquí considerado es el de medida. Lo que parece obvio es que la calificación es el resultado de un examen (observación, comprobación, interrogación, prueba objetiva, etc.) y en este sentido, exige medir. Sin medición no es posible hablar de calificación mínimamente objetiva y fiable, y por ende, no es posible hablar tampoco de evaluación.

Los tres conceptos, pues, evaluación, calificación y medida, están íntimamente ligados. Ciertamente, el valor de la evaluación y de la calificación (objetividad, fiabilidad o validez) dependen de los criterios de examen y estimación; de aquí que existan distintas formas o tipos de calificación, dependiendo del uso a que estén ordenadas, del país, del sistema escolar, del nivel de enseñanza y de otros elementos. De igual forma, existen también distintas formas o tipos de evaluación, que analizaremos más adelante.

Siguiendo con el análisis del concepto de evaluación, y además de ser en su núcleo conceptual un juicio de valor sobre una realidad, la evaluación se caracteriza además por su carácter instrumental, esto es, evaluamos algo que trasciende a la propia evaluación. Se evalúa para tomar decisiones en el contexto educativo, para mejorar la enseñanza y el aprendizaje. Así pues, como afirma De la Orden (1982), "al juicio de valor en que la evaluación consiste, sigue una decisión... Ciertamente que en ocasiones la mera formulación del juicio puede parecer la única decisión: tal alumno ha comprendido el concepto; tal profesor es eficiente; tal departamento funciona bien; tal centro es mejor que tal otro, etc. Pero aun en estos casos, hay decisiones implícitas que van más allá del propio juicio de valor: este alumno ha comprendido, luego no necesito repetir la explicación; como tal profesor o departamento o centro son eficientes, me quedo satisfecho, etc. Estas consecuencias y otras de tipo similar constituyen la manifestación mínima de la toma de decisiones inherente a todo proceso educativo" (págs. 17-19). El diagrama que presentamos a continuación - tomado de De la Orden (1982, pág. 18) - presenta en síntesis una estructura de la evaluación educativa coherente con las características antes señaladas como definitorias del concepto de evaluación (figura 2.1).

•En el contexto educativo, como vemos, pueden evaluarse diferentes instancias de la realidad educativa:

-El aprendizaje de los alumnos.

-La eficacia del profesor.

-La eficacia de un programa.

-La eficacia diferencial de diferentes técnicas o métodos didácticos.

-La eficacia relativa de diferentes materiales didácticos.

-La estructura y organización de un departamento o de un centro educativo.

-La eficiencia de un esquema de evaluación.

-Etc.

De hecho, la evaluación educativa debe abarcar todo el conjunto de factores significativos que inciden, no sólo en el aprendizaje, sino también en la educación integral de los sujetos. Pero a la vez, debe referirse también a todo el conjunto de elementos que forman parte del proceso educativo, condicionando o facilitando la mejora de los educandos.

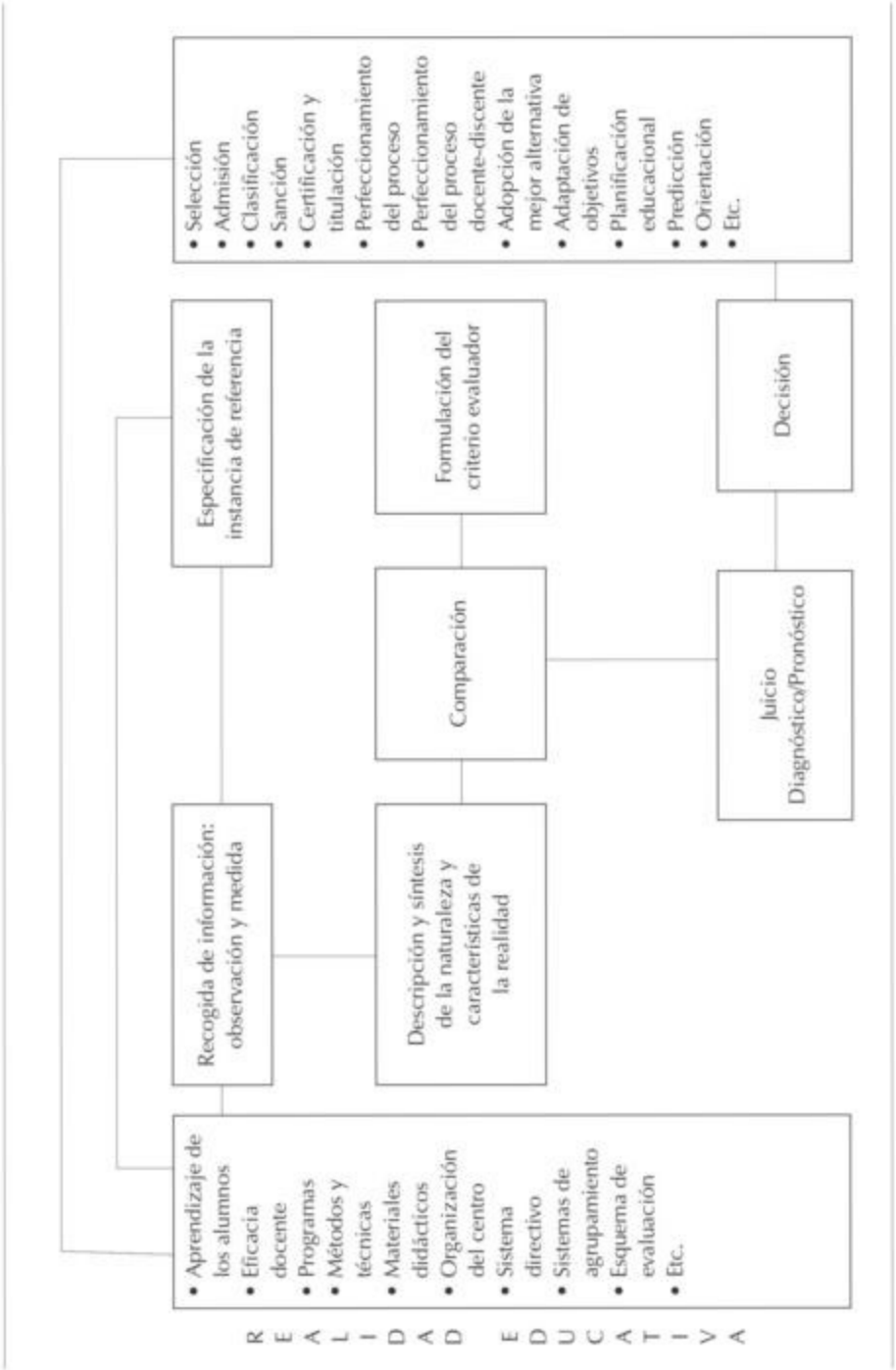


Figura 2.1. Modelo de Evaluación (De la Orden, 1982).

En este trabajo nos vamos a ocupar fundamentalmente de la evaluación del aprendizaje de los alumnos, esto es, de las características, tipos, modelos y funciones de la evaluación de los conocimientos adquiridos por un alumno en el contexto de una institución educativa, aun cuando sea preciso hacer referencias ocasionales a la evaluación de otros contextos educativos.

•Quizás el punto de partida de este análisis sea la tesis del profesor De la Orden (1969,

1981, 1982) de que la evaluación, sea cual sea el sistema utilizado, tiene una influencia decisiva y directa sobre la calidad de la educación. Esta tesis, que dicho autor formula así: "sea cual fuera la modalidad, sistema o tipo de evaluación vigente en un contexto educacional, su influencia sobre el proceso y el producto de la educación es decisiva y afecta directa o indirectamente a su calidad" (pág. 23), nos debe invitar a reflexionar sobre la importancia que debe darse al sistema evaluativo utilizado, y a la evaluación educativa en general, dado que, según este planteamiento, el modelo de comportamiento implícito en los criterios de evaluación - al constituir una expresión de lo que se espera del alumno, puesto que es lo que realmente se le exige-, define de una manera efectiva los objetivos reales y operantes de la educación, con independencia de lo formulado en los programas y de que existan o no formulaciones expresas o no de los mismos.

En consecuencia y así contemplada, la evaluación se constituye en motor y palanca del aprendizaje del alumno y

> "determina en gran medida - al prescribir realmente los objetivos de la educación - las características de la enseñanza y del aprendizaje, lo que los alumnos aprenden y cómo lo aprenden, lo que los profesores enseñan y cómo lo enseñan, los contenidos y los métodos; en otras palabras, el producto y el proceso de la educación" (De la Orden, 1982, pág. 25).

Dada la importancia que concedemos a este enfoque, continuemos con el análisis del mismo autor:

> "en efecto, si la calidad de la educación está vinculada a un sistema de coherencia con el proceso total de la educación, la virtualidad de la evaluación, como estímulo o freno de la calidad educativa, dependerá enteramente de las características del esquema evaluador. Si los criterios y modos de evaluación, que determinan los objetivos reales y, en última instancia, el producto de la enseñanza y el aprendizaje son coherentes con los objetivos formalmente establecidos - o sobre los que existe un consenso implícito - y a través de ellos con los fines generales de la educación y el sistema de valores del que derivan, la evaluación actuará como el más poderoso factor de promoción de la calidad educativa, al garantizar la congruencia y eficacia del sistema en su conjunto. Si, por el contrario, los criterios y modos de evaluación movilizan los procesos educativos hacia objetivos y resultados no coherentes con las metas formalmente establecidas o implícitamente aceptadas, ni en consecuencia con los fines generales de la educación y el sistema axiológico en que se sustentan, la evaluación se constituye en el mayor obstáculo a una educación de calidad" (págs. 25-26).

•Quizás sea también importante diferenciar los conceptos de evaluación e

investigación. Es evidente que están íntimamente relacionados, dado que en el proceso de evaluación y en el de investigación existen muchos elementos comunes, tales como: la necesidad de formular los objetivos, la necesidad de especificar variables, la importancia de diseñar el proceso, la necesidad de medir, herramientas metodológicas de análisis, etc.; ahora bien, las diferencias entre ambos conceptos son claras y básicamente radican en la finalidad diferencial de ambos procesos. Así, mientras que la evaluación es un proceso que busca información para la valoración inmediata y la toma de decisiones sobre el valor de algo, la investigación es más bien un proceso que busca principios y leyes, conocimiento generalizable (al menos la investigación básica), un proceso que busca conclusiones. Quizás, en ocasiones y en el intento de unir investigación y praxis educativa, se han aproximado los conceptos de investigación y evaluación. Lo que debe quedar claro es que la investigación busca fundamentalmente conclusiones para la construcción de teoría, sobre la base de distintos modelos o paradigmas, pero sin que dicha teoría tenga necesariamente una aplicación inmediata, puede tenerla o no tenerla. La evaluación se centra en un fenómeno particular. No pretende generalizar los resultados a otras situaciones. Su interés está en esa situación y en las decisiones a tomar sobre ella.

- Así pues, la evaluación es una actividad o proceso de identificación, recogida y tratamiento de datos sobre elementos y hechos educativos con el objetivo de valorarlos primero y, sobre dicha valoración, tomar decisiones. En palabras de Pérez Juste (1986),

> "evaluar es el acto de valorar una realidad, formando parte de un proceso cuyos momentos previos son los de fijación de las características de la realidad a valorar, y de recogida de información sobre las mismas, y cuyas etapas posteriores son la información y la toma de decisiones en función del juicio de valor emitido" (pág. 31).

Pophan (1980) utiliza una definición más práctica aún: "la evaluación educativa sistemática consiste en un juicio formal del valor de los fenómenos educativos" (pág. 16). En esta definición, destacan los conceptos sistemática (para diferenciar a la evaluación educativa de la evaluación diaria informal) y formal (en el sentido de determinación del mérito de una actividad, concepto éste sobre el que Scriven (1976) ha insistido reiteradamente al afirmar que "la valoración del mérito es la esencia del acto evaluativo").

Así pues, y como síntesis, la evaluación implica tanto el procedimiento de recogida de información como el procedimiento de toma de decisiones. No puede, de hecho, hablarse de evaluación si falta la referencia a alguno de estos procedimientos. El primero implica medir, el segundo valorar, decidir.

Quizás la definición del Comité Phi Delta Kappa de Evaluación de la Enseñanza Nacional (1971) es de las que mejor expresan el concepto y el contenido de la

evaluación: "el procedimiento que define, obtiene y ofrece información útil para juzgar decisiones alternativas", esto es, para decidir cuál, de entre las opciones disponibles, es la más adecuada y útil para alcanzar unos objetivos o para valorar si unos objetivos se han cumplido o no y en qué grado.

2.1. La evaluación y su evolución conceptual

El concepto de evaluación ha evolucionado a lo largo de los últimos cincuenta años desde la conceptualización de Tyler hasta nuestros días.

Actualmente y como consecuencia de su desarrollo histórico, el concepto de evaluación se ha ampliado enormemente.

Toda actividad educativa intencional es susceptible de evaluación y por tanto es evaluación educativa.

El término inglés evaluation es un término mucho más amplio, referido a la estimación sobre el valor o la efectividad/ eficacia de programas de intervención o el análisis de resultados de actividades educativas. En este sentido, puede decirse que es un término que de alguna manera engloba al otro término (Assessment).

En este sentido, la diferencia básica entre ambos conceptos parece estar en la unidadobjetivo de análisis: un individuo o sujeto único (assessment) o bien, un programa o modalidad de intervención (evaluation).

Si tuviéramos que elaborar una definición personal, no se alejaría demasiado de otras ya dadas con anterioridad (De la Orden, 1982; Pérez Juste, 1986; Mateo, 1986...):

> Una actividad o proceso de identificación, recogida y tratamiento de datos (fiables, válidos y útiles) sobre elementos y hechos educativos, con el objeto de valorarlos primero y, sobre dicha valoración, tomar decisiones.

En esta definición, se asumen también otras características que toda evaluación debe mostrar: sistematicidad, rigor, precisión, contexto definido y referencia al objeto o unidad de análisis (individuo o programa).

La actividad humana tiene presente la necesidad de Evaluación intencional y sistemática.

Tradicionalmente la evaluación se consideró la tercera fase del ciclo completo de instrucción (programación, metodología y evaluación).

Hoy en día no se limita a la constatación de una fase terminal del aprendizaje de un sujeto. A ello han contribuido:

-La introducción del concepto de valor.

-La necesidad de comprobar la eficacia y éxito de infinidad de programas.

La terminología "Evaluación de Programa" surgió con el objetivo mismo de evaluar programas de acción que se venían aplicando y sobre los cuales no había suficientes indicios objetivos de su eficacia.

El método concreto de evaluación se ha venido denominando "investigación evaluativa"".

La investigación evaluativa es una manera de aumentar la racionalidad de las decisiones.

•¿Cuáles son los hitos fundamentales de la breve historia del concepto de evaluación? Siguiendo a Carballo (1991):

1.Se retomó, a partir de 1960, el modelo de Tyler cuyos elementos esenciales son: objetivos, o conductas deseables, Actividades de aprendizajes (individuales o de grupo), y calificación de los aprendizajes (rendimiento).

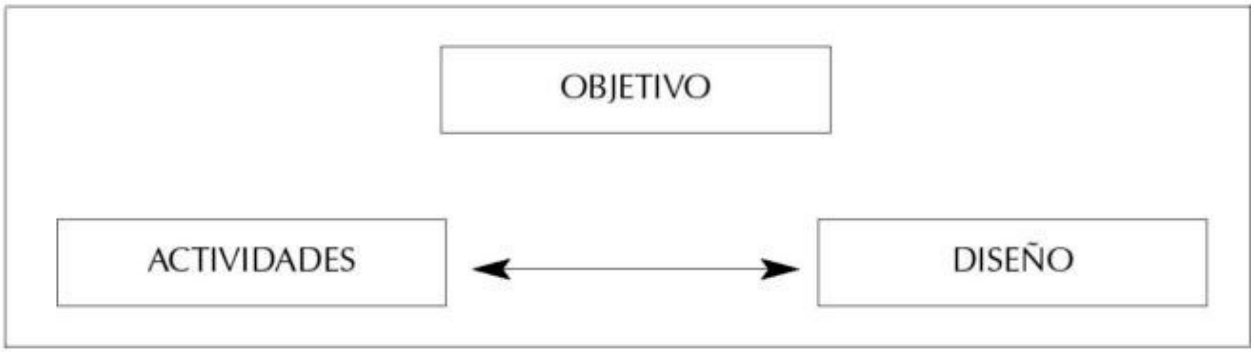


Figura 2.2. Modelo de Tyler.

"Evaluación como proceso para determinar si se han alcanzado los objetivos y de acuerdo con esto se decide si el programa es eficaz, ha tenido éxito".

(+)Gran influencia del modelo de Tyler; otros modelos abreviados:

-EPIC de Hammond (1972).

-Discrepancias de Provus (1971).

-Taxonomías de Bloom (1956).

-Objetivos instructivos de Popham (1969).

-Modelo de ocho etapas de Metfessel y Michael (1967).

(-)Críticas:

-Grandes lagunas en los componentes evaluativos (valen para medir y valorar objetivos, más que para enjuiciar el éxito global de un programa).

-Si no alcanza los objetivos, ¿puede ser eficaz el Programa?: Sí.

-Si se alcanzan los objetivos, puede ser ineficaz el Programa?: Sí.

2.Esta limitación última motivó que se pusiera en entredicho la validez de un programa atendiendo sólo a los objetivos, reflejados en las calificaciones de los alumnos.

A finales de los años 60, los primeros evaluadores incorporados a esta tarea procedían de la investigación y se generalizó el auge de los diseños experimentales, cuas ¡experimentales y causales.

Evaluación como investigación para demostrar la eficacia de un programa a través de la relación causal entre objetivos y programa.

(Modelos de Evaluación basados en la experimentación).

Numerosos estudios:

-Linquist (1953).

-Campbell y Stanley (1966).

-Suchman (1967).

-Wiley y Bock (1967).

-Gloss y Maquiere (1968).

-Cronbach y Snow (1969).

-Etc.

Surgieron dos tendencias:

a)Defensores del experimento puro.

b)Menos estructuración y mayor flexibilidad en los experimentos.

Se potenció la concepción comparativa de la evaluación:

•Comparación con objetivos.

•Comparación con un ideal.

•Comparación con otro modelo.

(+)Han desempeñado un papel muy relevante, han proporcionado: métodos válidos y consistentes para determinar relaciones causales relativamente inequívocas entre un programa y sus resultados.

(-)Críticas: (limitaciones propias del método experimental).

-No facilita toda la información necesaria para tomar decisiones.

-Programas no comparables en muchas ocasiones.

-(Objetivos distintos o no equivalentes).

3.Ello provocó la necesidad de considerar la evaluación de los mismos objetivos de los programas, para ver si: eran correctos o erróneos, utópicos o factibles, y relevantes o irrelevantes, respecto de las necesidades para las que surgió el programa.

•Se modificó el esquema de Tyler y surgieron los modelos orientados al juicio.

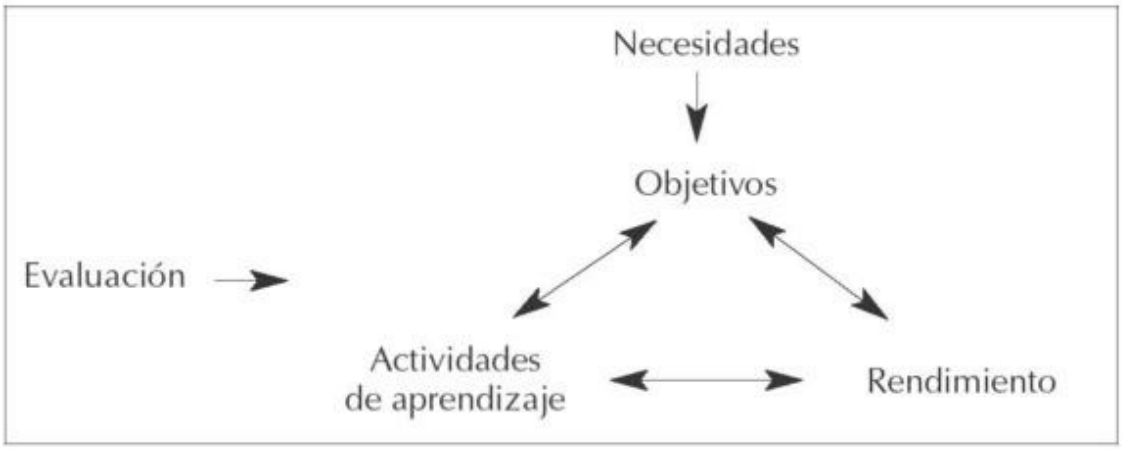


Figura 2.3. Modelo orientado al juicio.

Se basaban en la competencia valorativa y profesional del evaluador para emitir juicios. Su función, por definición, era "científica" y, por consiguiente, "objetiva".

La realidad demostró que se basaba, en criterios valorativos previamente seleccionados y justificados, por lo que la pretendida "objetividad" desacreditó los estudios evaluativos mismos.

Scriven defendió el concepto de valoración como una característica "sine qua non" de toda evaluación, y que evaluar significaba valorar: objetivos, resultados y decisiones.

Surgieron así los modelos de:

-Stake (1967) "de la Figura".

-Eisner (1971) "consulta a experto".

Estos modelos asignaban al evaluador el papel de `crítico educativo"

(+)Papel de la valoración y juicio de las personas, y la importancia de los criterios elegidos para valorar.

(-)Modelos que llevan a dar juicios utópicos al evaluador.

Competencia del evaluador (peor debilidad).

Los evaluadores se sentían pertenecientes a una "selecta minoría" (incomodidad).

4.Como alternativa a estas críticas, surgieron los modelos políticos o de debate, que incluyen evaluadores de distintas perspectivas y procedimientos jurídicos y de debate político y técnico, adaptados a la evaluación de programas.

•Modelo contrapuesto de Owens (1973).

•Modelo judicial de Wolf (1974, 1985).

•Modelos implícitos en estudios de: Clark (1963), Coleman y otros (1966), Jenks y otros (1972), Weis (1972).

(+)Servían como guía de las instituciones y de la sociedad.

(-)Acaban corrompidos (los modelos) o viciados por el ambiente político en el que se desarrollan.

5.Surgieron entonces los modelos de evaluación de costos, cuya finalidad era determinar las necesidades de inversión económica en cada programa, y elegir el más barato (y de más calidad).

Sin embargo, la relación calidad-costo se descubrió como muy importante. A mayor costo, mayor calidad, mejores aprendizajes, etc.

Se hicieron estudios basados en:

•Costo - Beneficio.

•Costo - Eficacia.

•Costo - Utilidad.

•Costo - Viabilidad.

Todos los estudios permitían valorar si las inversiones efectuadas son adecuadas y rentables (Woodhall, 1967; Levin, 1975, 1983, 1977; UNESCO, 1972, 1980, etc.).

6.En esta época, Scriven muestra la necesidad de valorar los programas atendiendo, no sólo a los objetivos propuestos, sino también a las consecuencias no previstas, que pueden derivarse de su aplicación (evaluación libre de objetivos) ("goal free evaluation").

El que la evaluación sea soporte para la toma de decisiones administrativas, políticas y, por supuesto, educativas, es una característica señalada en general por todos los autores.

Y, aunque la mayoría de las teorías de la evaluación manifiestan su intención de que sus contribuciones y modelos influyan y ayuden a sus clientes, en la mayoría de los casos esta influencia no ha existido o ha sido poco significativa:

•Evaluación centrada en la utilización.

•Evaluación orientada a la toma de decisiones.

Dos posturas se han adoptado ante el problema del uso de las evaluaciones para la toma de decisiones:

-Centrarse en las necesidades de los implicados en la toma de decisiones.

-Comprometerse el evaluador y la propia evaluación en incrementar y mejorar los programas.

La primera postura dio lugar a modelos como los siguientes:

•Planificación educativa a través de "UTOS" (Cronbach, 1963 y 1982).

•Modelo CIPP (Stufflebean, 1966 y 1967).

•Modelo centrado en el cliente (Scriven, 1967).

•Modelo de evaluación formativa (Brinkerhoff y otros, 1983).

•Estudios de:

-Alkin (1969)

-Reinhard (1972)

-Taylor (1984)

-Guba (1978)

-Webster (1975)

-Etc.

En este caso, la eficacia de dichas evaluaciones depende mucho de la coordinación del trabajo en equipo del evaluador con los encargados de la toma de decisiones.

(+)Estas estrategias animan a utilizar la evaluación de una forma continua y sistemática, y a poner a disposición de los usuarios los servicios que satisfagan sus necesidades.

(+)Proporciona además, a los profesionales, una base lógica que permite responder a las decisiones adoptadas durante el curso del programa.

(-)Se les critica que con frecuencia, no son capaces de determinar de manera explícita el éxito de un programa.

(-)Dependen demasiado de la colaboración entre Evaluador y Responsable de la toma de decisiones, colaboración que no suele ajustarse a supuestos ideales, lo que puede distorsionar los resultados.

7.La segunda postura, reivindicó más el compromiso del evaluador y de la evaluación, con la mejora de los programas.

De este modo, son las valoraciones y los puntos de vista de aquel que es evaluado, las que tienen más peso (evaluación pluralista o iluminativa).

Ejemplos de esta postura son:

•La evaluación respondente de Stake (1975).

•La evaluación iluminativa de Parlett y Hamilton (1977).

•La evaluación transaccional de Rippey (1973).

•La evaluación democrática de McDonald (1975).

(+)Conducen a investigaciones activas, de gran riqueza informativa, que proporcionan a todas las personas que participan en el programa los recursos necesarios para dirigir sus propias evaluaciones.

(-)Su defecto típico radica en la ausencia de credibilidad externa, y en la posibilidad de que sean manipuladas por determinados elementos locales que ejercen mayor control sobre la evaluación.

8. En ambas posturas, (7) y (8), se ven reflejadas las dos funciones que según Scriven (1967) desempeña la evaluación:

-Sumativa (7): Ayudar al tomador de decisiones sobre el valor y la continuidad del programa, una vez finalizado.

-Formativa (8): Servir de ayuda a las personas involucradas en el programa para perfeccionarlos mientras se desarrolla.

Todo ello llevó a Scheerens (1987) a modificar el esquema anterior planteado por Wolf, añadiendo la fase de toma de decisiones. En este esquema aparecen claramente diferenciadas las dos posibilidades de la evaluación.

Figura 2.4. Modelo de evaluación (Scheerens, 1987).

En el caso de la primera, la evaluación es parte integrante del proceso educativo y en el caso de la segunda, la evaluación está orientada a la toma de decisiones sumativa, una vez que se ha comprobado la eficacia del programa respecto a las necesidades planteadas, y después de haber llevado a cabo las mejoras oportunas.

2.2. Evolución del concepto de evaluación

- Los diferentes modelos de evaluación que surgen y la práctica evaluativa siguen aportando elementos para esta definición y van encontrando nuevas soluciones y nuevos problemas.

- En la actualidad, es la práctica evaluativa la que está aportando los mayores frutos, a través de: manuales y guías prácticas para realizar evaluaciones, la elaboración de nuevos instrumentos evaluativos, creación de gabinetes de evaluación, etc.

- Quizás, como dicen Cooky Shadish (1986), va llegando el momento de reorganizar y sistematizar la información existente; para ello, estos autores proponen cuatro bases teóricas de reestructuración de lo ya aparecido hasta el momento:

1. La Teoría de los programas sociales, que permitirá al evaluador saber cómo poder

producir cambios sociales (en nuestro caso, en el campo educativo), y con qué limitaciones se encuentra.

2. La utilización de la evaluación, que proporcionará al evaluador información de cómo se aplica el conocimiento de las ciencias sociales a la programación social.

3. La Teoría del valor, que pondrá a disposición del evaluados datos que le ayuden a diferenciar los buenos programas de los malos.

4. Las bases metodológicas, o de construcción de conocimientos válidos, que facilitará al evaluador el camino para llegar a conclusiones válidas.

Y una quinta base que yo añadiría:

5. Una teoría de la práctica evaluativa, que sirva para guiar y hacer explícita y racional la labor del evaluador.

2.3. Algunas conclusiones sobre la evaluación y su incidencia en la calidad educativa

Visto lo anterior, parece claro que la evaluación tiene un carácter instrumental, ya que se evalúa para tomar decisiones que mejoren la enseñanza, el aprendizaje, la puesta en marcha de un programa educativo..., en definitiva, la calidad de un proceso educativo definido en orden a la consecución de unas metas u objetivos.

Algunas conclusiones que me parecen importantes referidas a la evaluación podrían ser las siguientes:

- Si no hay metas y objetivos educativos claros y definidos, que orienten la actividad educativa, la evaluación es una herramienta inútil e innecesaria.

- Si no está orientada a la toma de decisiones, aportando como una fuente más su valoración de los distintos aspectos evaluados, la evaluación no conduce a ningún sitio, convirtiéndose en un ejercicio en sí mismo sin utilidad alguna.

- Si las decisiones ya están tomadas y no se puede o no se quiere cambiar la orientación de un programa o acción educativa, la evaluación es un pasatiempo inútil y, "como todo esfuerzo inútil, conduce a la melancolía del equipo evaluador y coordinador".

- Esperar una claridad total de los resultados de un proceso educativo equivale a mecanizar la enseñanza y el aprendizaje y, en parte, "cosificar" a la persona o programa evaluado.

La evaluación debe entenderse como una entrada más (muy importante, pues objetiva y valora lo ocurrido) al proceso "político" de toma de decisiones, y nunca como la única

entrada o la más importante.

Muchos de los desencantos que se han dado en el uso e investigación sobre enseñanza, se deben a las excesivas expectativas que sobre la evaluación se pusieron.

Finalmente, me gustaría terminar con la tesis fundamental que sobre la evaluación, como instrumento, podemos contemplar y que ya desarrollamos al principio de este capítulo.

Sea cual sea el sistema utilizado, la evaluación tiene una influencia decisiva y directa sobre la calidad de la educación.

La tarea educativa es exigente y con frecuencia ingrata. Por ello, no será fácil encontrar quien lidere y se comprometa de manera real en la labor de dirección, ejerciendo un liderazgo que encamine a la comunidad educativa hacia los objetivos marcados.

3

El rendimiento educativo. Acotaciones

Cuando se pretende estudiar la eficacia de un programa educativo o de un centro, o bien el aprendizaje de un alumno como efecto del proceso de enseñanza-aprendizaje en el que está inserto, es obvio que necesitamos criterios, indicadores que nos permitan formular juicios o tomar decisiones.

Por lo general, los criterios base que se vienen utilizando como indicadores de la eficacia son medidas del producto de la educación. Si estamos interesados en evaluar, no sólo los conocimientos y destrezas adquiridas por un sujeto, sino también la forma de adquisición de tales conocimientos y destrezas, las actitudes generadas durante el aprendizaje e incluso la significatividad de los avances conseguidos por sus alumnos en un área concreta, con un punto de partida conocido, nos interesa estudiar no sólo variables de producto, sino también variables de proceso y variables antecedentes al propio proceso de enseñanza-aprendizaje que se pretende evaluar.

Esta clasificación anterior de las variables que interesa estudiar y analizar desde el punto de vista de la evaluación (variables antecedentes, variables de proceso y variables de producto) sin duda tiene que ver con una de las clasificaciones más comunes de la evaluación: evaluaciones inicial, continua y final.

3.1. ¿Qué es el rendimiento?

No obstante lo anterior, en este momento es preciso hacer algunas acotaciones sobre lo que viene comúnmente llamándose rendimiento educativo, al objeto de situar la evaluación de los conocimientos y aprendizajes de un alumno en un contexto concreto.

Ciertamente, la combinación de todos los factores o variables implicadas en los procesos de enseñanza-aprendizaje hacen que el comportamiento educativo - el comportamiento como efecto de la educación - sea diferenciado. Lógicamente, es el llamado Rendimiento Educativo o producto de la educación, el que sufre la influencia de todas esas variables, de ahí que el rendimiento sea también diferenciado, es decir, varíe entre sujetos.

Ahora bien, ¿qué es el rendimiento educativo?, ¿qué queremos medir como efecto de la educación?, ¿en qué grado las medidas del rendimiento expresan bien el resultado del

proceso educativo?, ¿qué entendemos por producto educativo? Todas estas cuestiones nos conducen a la necesidad de conceptualizar y definir el producto de la educación. De hecho, el tema del rendimiento es capital en la evaluación de cualquier proceso educativo y es de vital importancia en el avance de la ciencia pedagógica a partir de la investigación educativa.

En principio y a primera vista, como afirma De la Orden (1985), el producto de la educación podría definirse como "el resultado de una acción o de un proceso educativo". Ahora bien, dicha definición es ambigua, no operativa, excesivamente general y limitada en su concepción individual, al no considerar la proyección social de los efectos de la educación. En este sentido, cabe el peligro de pensar que el producto educativo es "aquello que se mide en las escuelas". Ciertamente, la escuela tiene como objetivo básico el incremento y la evaluación de los procesos de aprendizaje en términos de rendimiento educativo, pero, ¿de qué tipo de rendimiento hablamos?

Es obvio, pues, como afirma De la Orden (1985), que "necesitemos una Teoría del Producto Educativo que guíe la Acción Educativa". Hoy por hoy, no la tenemos, pero sin duda debemos buscarla, especialmente si queremos avanzar en el conocimiento de los efectos de la Acción Educativa y de cómo mejorar dicha acción.

Actualmente sólo tenemos algunas teorías operacionales que cubren una parte de lo que podríamos llamar producto de la educación. Son las implicadas en las taxonomías de objetivos educacionales (Gagné, Bloom, etc.). Estas taxonomías sólo cubren lo que podríamos denominar producto individual inmediato y únicamente en la esfera del denominado dominio cognitivo (conocimientos, comprensión, aplicación, etc.).

Si como dijimos anteriormente, el producto de la educación es "lo que acontece al estudiante en el proceso educativo", es lógico preguntarse ¿qué le acontece? La realidad es que a esta cuestión no hay respuesta empírica clara. Tenemos sólo algunas respuestas concretas que no suelen ser expresión de capacidades generales. Lo que es evidente es que no estamos interesados únicamente en el resultado individual inmediato del proceso educativo, es más, en ocasiones estamos más interesados en el resultado mediato individual y en los resultados sociales de la acción educativa.

Con estas dos dimensiones (individual-social y mediato-inmediato), De la Orden (1985) define cuatro tipos de productos educativos, como resultado del cruce o combinación de los niveles o categorías de ambas dimensiones (figura 3.1).

Individual	Social
Inmediato individual	Inmediato social
Mediato individual	Mediato social

Figura 3.1. Tipos de producto educativo.

Dentro de cada cuadro podríamos, a su vez, realizar otras divisiones (por ejemplo, cognitivo-no cognitivo, etc.).

La siguiente figura (figura 3.2), tomada del mismo autor, puede darnos una idea más clara de los diferentes tipos de producto educativo categorizables desde una visión amplia del producto educativo.

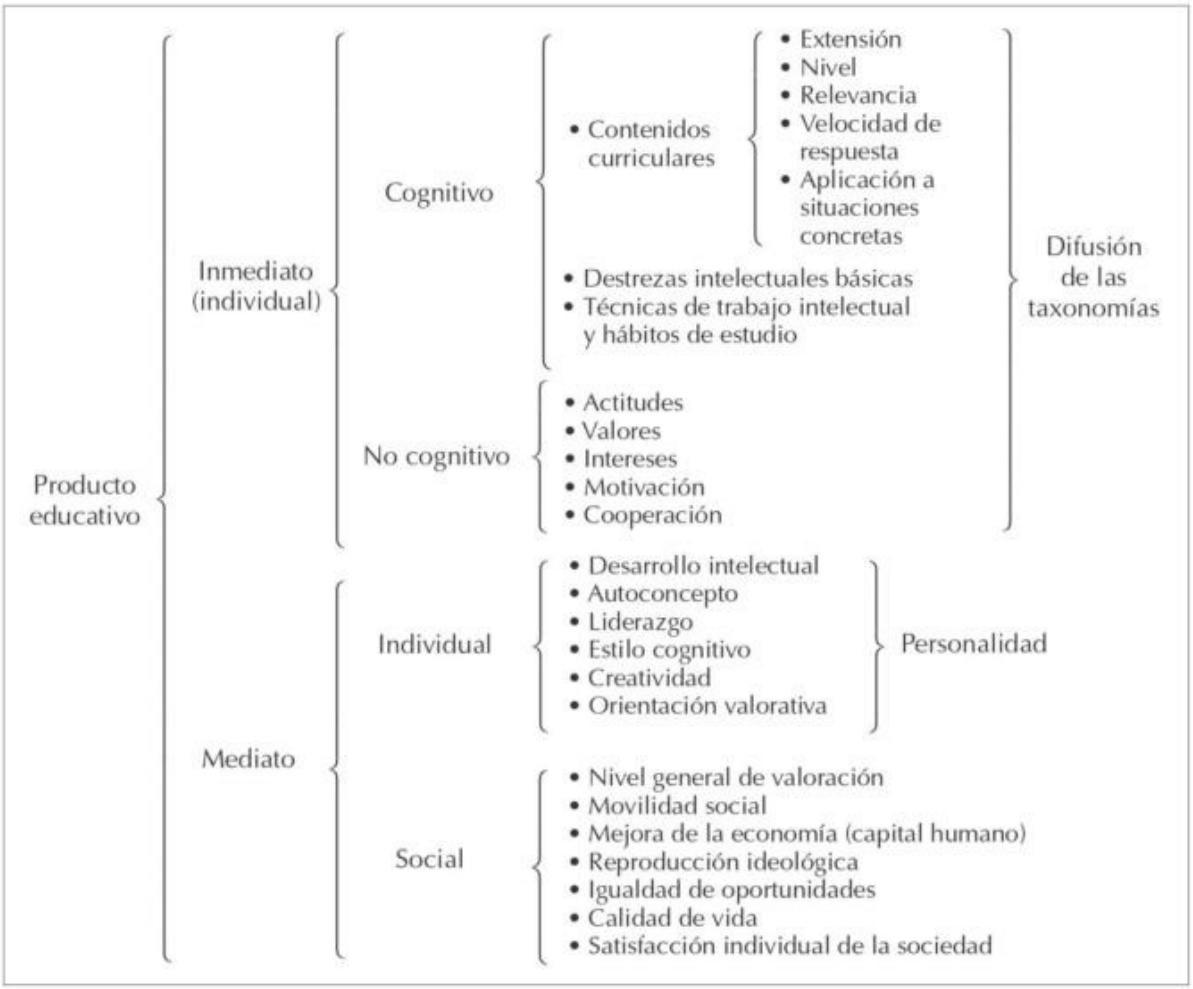


Figura 3.2. Clasificación de los productos educativos.

La evaluación del rendimiento, a la que comúnmente nos solemos referir, es la evaluación de los conocimientos, o mejor dicho, la evaluación referida al producto individual inmediato de la educación. Es en este contexto en el que nos movemos básicamente en este libro, pero sin olvidarnos de que existen otros tipos de productos educativos - ciertamente menos operativos, más difíciles de precisar y de medir - que en la mayoría de las ocasiones son objetivos finales más importantes de la escuela y de la propia educación.

Lo que sí parece claro es que no hay criterios unívocos específicos y generalizables para medir el rendimiento, el cual se presenta como una realidad evasiva y multidimensional, y ello porque no hay - como afirma De la Orden (1985) - una teoría clara del producto educativo que permita la definición de criterios generalizables de medida. Por ello, pese a todos los empeños y esfuerzos, hoy por hoy, las calificaciones y certificados escolares y académicos son muy subjetivos, poco fiables y de escaso valor pronóstico y predictivo.

Como índices tradicionales del producto educativo se han venido utilizando pruebas de aptitud pedagógica (y de rendimiento), las cuales tienen por base la competencia demostrada por alumnos en las mismas. Hasta hace pocos años, no había más selección posible que pruebas referidas a normas. Con la aparición de mediciones que hacen referencia a un criterio, son sustanciales las diferencias introducidas. En un apartado posterior se analizará el tema al mencionar las diferencias entre la evaluación de referencia normativa y la evaluación de referencial criterial.

3.2. Evaluación del rendimiento: formas e instrumentos

Para la evaluación del rendimiento - entendido como producto individual inmediato - el sistema más extendido es el examen a través de pruebas escritas u orales que se traducen en una escala previamente establecida, formada por un número de categorías que oscila entre dos y siete (si la escala es cualitativa) o entre cero y diez, o entre uno y seis (si la escala es numérica). Si la escala es cualitativa, suelen utilizarse categorías como: matrícula de honor, sobresaliente, notable, bien, aprobado, insuficiente, etc. Las escalas binarias (apto-no apto) sólo se utilizarán en sistemas selectivos, aunque una traducción de esta escala binaria se utiliza actualmente al definirse las categorías: "Necesita mejorar", "Progresa adecuadamente", en el sistema evaluativo del ciclo inicial de la enseñanza primaria.

En las escalas numéricas, se asigna a los sujetos una cantidad, dentro de una escala que corresponde con categorías como: muy deficiente, deficiente, suficiente, bien, notable y sobresaliente (en la escala de uno a seis, por ejemplo). En realidad estos valores no son más que números cualificados que no se diferencian de las palabras más que en la posibilidad de operar aritméticamente.

La validez de estos juicios depende en gran medida de la formación y experiencia del que juzga y, por ello, se presta a errores de origen diverso ("efecto halo", tendencia a la severidad, a la benignidad o a evitar los extremos).

> "El análisis de las calificaciones confirma empíricamente que los enjuiciamientos resultan sistemáticamente diferentes en las diversas asignaturas, escuelas, tipos de escuelas, grados escolares y según el sexo" (Braum y Klink, 1983).

Las calificaciones escolares sobre la base de medidas objetivas del rendimiento (tests informales o tests escolares estandarizados) reducen el margen de error y elevan el grado de objetividad y fiabilidad (los analizaremos más adelante).

Como sistema de puntuación y calificación, en ocasiones se utilizan escalas técnicas que ofrecen ciertas ventajas de comparación entre sujetos y entre grupos, sobre las puntuacio nes directas o brutas; algunos de ellos son: centiles, puntuaciones típicas, puntuaciones derivadas, etc.

Por otra parte, para juzgar el rendimiento hay núcleos de calificación que tienen en cuenta objetivos de aprendizaje y los diversos grupos de sujetos. Los modelos con función de objetivos (de referencia criterial) miden el rendimiento del individuo en proporción a las metas fijadas; los modelos en función del grupo (de referencia normativa) lo miden con relación a un grupo de referencia, que puede ser una clase concreta, un grupo particular o un grupo representativo tomado como punto de comparación. En la praxis espontánea, sin embargo, las calificaciones, por lo general, se atienen a los dos modelos, sin distinguirlos de forma alguna.

Normalmente, las calificaciones escolares o académicas sirven para orientar al alumno en la elección de la formación subsiguiente, esto es, tienen valor pronóstico o predictivo de ulteriores rendimientos académicos, cuya fiabilidad se limita a juicios estimativos o probabilísticos. Cuanto a más largo plazo quiera hacerse la predicción, tanto más limitada será la estimación, dado que, como puede comprobarse empíricamente, las correlaciones entre predictor (variable medida, base para la predicción) y criterio (variable a predecir) suelen ser tanto más bajas cuanto mayor sea el intervalo de tiempo entre la medida del predictor y el momento al cual se extiende la predicción del criterio. Ciertamente, es posible mejorar estas predicciones reduciendo las poblaciones en las que se quiere predecir (a escuelas, a especialidades, a asignaturas, etc.) o bien, incluyendo más variables predictoras del criterio, de distinta naturaleza al propio rendimiento (tales como capacidades generales y específicas, actitudes, intereses, de personalidad, etc.) en los modelos de regresión que son la base de los procesos predictivos.

En todo caso, y como anteriormente hemos señalado, las calificaciones escolares y los certificados, hoy por hoy, son muy subjetivos, poco fiables y de escaso valor pronóstico. Y ello, además de por no tener clara una teoría del producto educativo y por su realidad multiforme, porque en las calificaciones escolares intervienen muchos factores ajenos al rendimiento mismo, tales como: el tipo de escuela, el nivel de escolaridad, el sexo, etc., y por supuesto la persona del profesor. Un ejemplo claro puede ser el clásico efecto halo (tendencia a considerar que un alumno sobresaliente en una materia lo es en todas, o en todas es deficiente por serlo en una o varias) el cual influye a la hora de extender un certificado escolar.

Por otro lado, se discute la función estimulante de las calificaciones y se admite que constituyen un obstáculo a las motivaciones objetivas por el estudio de las asignaturas, mientras crean y refuerzan el espíritu de emulación y competencia (competitividad).

En su forma actual, la mayoría de los centros coinciden en admitir que tanto calificaciones como certificados escolares - como apreciaciones del rendimiento escolar - no pueden aceptarse como datos académicos suficientemente comprobados y apoyados sobre las capacidades, conocimientos, asimilación, intereses, trabajo y aprendizaje de los alumnos. Hay deseos e intentos de reformar el sistema; se habla, por ejemplo, de "eliminar la evaluación numérica o completarla con un acta explicativa verbal", se piensa

en "diagramas de rendimiento"; algunos autores piensan que se deben intensificar los esfuerzos por "reducir al mínimo imprescindible" la composición de las notas, al objeto de simplificar el proceso, dentro del mínimo margen de error (puede hacerse con estudios de análisis factorial); sin embargo, otros autores defienden posturas contrarias y estiman que hay que "objetivar al máximo la apreciación del rendimiento mediante sistemas complejos de puntuación" (para ello, proponen el uso de curvas diagnósticas individuales que permitan describir objetivamente el rendimiento y orientar individualmente a los alumnos).

En todo caso y hasta ahora, parece que las esperanzas son mayores que los hallazgos. Es evidente que será necesario seguir buscando sistemas fiables y objetivos de calificación y evaluación, sistemas fiables y precisos basados en la medición y estudio de variables del producto educativo. Insistiremos en que, entre otras cosas, sería necesario que se intensificasen los esfuerzos por buscar teorías que permitan conceptualizar mejor lo que se denomina rendimiento o producto de la educación.

4

Modalidades de la evaluación

Aun cuando en las páginas anteriores se define la evaluación como un proceso valorativo de cara a la toma de decisiones, es evidente que técnicamente es necesario diferenciar distintas modalidades de este proceso atendiendo a diversos criterios clasificatorios.

4.1. Evaluaciones inicial, continua y final (momentos de la evaluación)

Si el criterio clasificatorio es el "cuándo" de su realización, las modalidades básicas de evaluación serían: inicial, continua o progresiva y final. Es evidente que, en el fondo, las tres modalidades podrían entenderse también como los pasos de un mismo proceso evaluativo global cuyo referente básico sería la evaluación progresiva o continua. En esta perspectiva, la evaluación inicial sería considerada como el primer paso de la evaluación progresiva, siendo la evaluación final, desde un punto de vista pedagógico, una resultante de la evaluación continua.

4.1.1. Evaluación inicial

Desde una perspectiva pedagógica es claro que el conocimiento previo del educando y de sus características, por parte del educador, es el punto de arranque de toda actividad educativa. Sin conocimiento de la individualidad del educando, difícilmente se podrá ayudar, objetiva y sistemáticamente, al desarrollo y formación de su personalidad. Este principio es básico en cualquier definición operativa de la educación, "nadie puede actuar sobre un material que desconoce". Por tanto, la evaluación inicial responde a este principio, el conocimiento concreto del ser real que debe educarse.

Ahora bien, ¿cuáles son las dimensiones del alumno que es necesario explorar de cara a una evaluación inicial? La respuesta puede ser doble, dependiendo de si el alumno inicia la escolaridad o se trata de la evaluación inicial de cada curso o si es referida a un cambio de centro del alumno dentro de un mismo curso. En ambos casos, la mayoría de los autores y la misma legislación española sobre evaluación coinciden en definir tres sectores o dimensiones exploratives, más una dimensión referida a los alumnos que no inician escolaridad:

a)Datos personales, ambientales y familiares.

b)Datos somético-sanitarios: datos médicos.

c)Datos psicológicos.

d)Antecedentes académicos.

Como vemos, tras la definición de las dimensiones de exploración, la evaluación inicial hace referencia a una situación estática del educando, a un corte horizontal de su trayectoria vital, mientras - como veremos - la evaluación continua o progresiva tiene un carácter dinámico y longitudinal.

No nos vamos a detener excesivamente en la descripción de estas dimensiones de exploración que pueden encontrarse en cualquier manual de evaluación con profusión de detalles, ejemplos e instrumentos, sin embargo, y a título orientativo, vamos a enumerar someramente los aspectos implícitos en cada una de las amplias dimensiones anteriores, estructurando el vasto campo de observación en los cinco apartados siguientes:

1.Ambiente familiar y social.

2.Constitución física "y salud.

3.Aptitudes o variables intelectuales.

4.Motivación, interés y actitud.

5.Adaptación o integración social: familiar, escolar y ambiental.

1. Ambiente familiar y social

Es evidente que la escuela no puede conformarse con una simple información de variables sociofamiliares del alumno, sino que debe buscar en todo momento que la formación de los educandos sea una tarea conjunta de la familia y la escuela, como punto de convergencia de ambas instituciones. Esta colaboración exige un recíproco conocimiento que ha de iniciarse en el momento del ingreso del niño en el centro educativo, para ir enriqueciéndose a lo largo de toda la escolaridad.

R.Sanmartín (1987), intentando sistematizar la información que el centro necesita para iniciar y desarrollar estas relaciones, estructura del siguiente modo la información a recoger (págs. 21-46):

A)Estructura familiar

a)Aspectos cualitativos

-Viven los padres.

-Huérfano de padre o madre.

-Viven en el hogar los abuelos.

-Viven tíos en el hogar.

-Empleada del hogar fija.

-Familia numerosa.

-Número de hermanos.

-Lugar que ocupa entre los hermanos.

-Hijo único.

b)Aspectos cualitativos (nota dominante de las relaciones familiares)

-Viven los padres juntos.

-Las relaciones son afectivas.

-Relaciones inarmónicas.

-Relaciones desordenadas.

-Relaciones rígidas.

-Padrastro o madrastra.

-Hijo tardío.

-Hijo natural.

-Hijo adoptivo.

-Niño privado de cariño.

-Educación emocional anormal.

-Niño educado por los abuelos.

-Niño confiado al cuidado de terceros.

-Padre fuera de la comunidad familiar (trabaja lejos del hogar).

-Madre que trabaja: medio día, todo el día.

c)Actitud familiar frente a problemas personales y escolares del niño

-Actitud superprotectora regresiva (niño mimado).

-Actitud superprotectora indulgente (niño consentido).

-Actitud estimulante (motivar al niño a que resuelva sus problemas personales y escolares).

-Actitud indiferente.

-Niño abandonado a sí mismo.

-Desaseo corporal y de la ropa.

-Vestimenta inadecuada a la estación.

-Asistir a la escuela sin haber desayunado o comido.

-Actitud negativa.

-Ambiente falto de estímulos.

-Incapacidad educativa de los padres.

-Educación inadecuada (pedagogía autoritaria).

-Oposición ala escuela.

B)Profesión u ocupación del padre (hace referencia directa a la vida escolar y futura vida profesional del niño)

Aquí no basta consignar tal o cual profesión. Es preciso "estimar" la dimensión social y "valorar" la fama de tal profesional y el grado de "satisfacción" o "insatisfacción" personal logrado en la profesión, puesto que ésta constituye, además de un medio para conseguir unos ingresos, una forma para llegar a una plena realización personal.

También es necesario conocer si la ocupación es indeseable personal o socialmente, subempleo, paro, etc. Si hay otros profesionales dentro del hogar - por ejemplo, hermanos - es conveniente consignarlo.

C)Nivel económico

-Conviene registrar aquí los ingresos familiares. Es un dato difícil de obtener con

fiabilidad, pero es importante.

-Dificultades económicas que alteran la serenidad familiar.

-Existencia de una vivienda capaz y estable.

-Vivienda sin servicios mínimos.

-Dormitorios compartidos por padres e hijos.

-Inseguridad en el disfrute de la vivienda.

D)Nivel cultural

Pueden consignarse varios indicadores. Uno referente a si se trata o no de un ambiente culto o no. Ello puede saberse conociendo los estudios realizados por los padres, sus títulos, etc., o los de los demás miembros de la familia (abuelos, hermanos mayores, etc.). El otro indicador es el relativo a la escala de valores cultivados en el seno familiar (económicos, estéticos, sociales, ideológicos, políticos o religiosos).

E)Ambiente social

Estudio de las relaciones sociales de la familia, las cuales aluden a la integración de la misma dentro de su propio contexto: muy relacionada, bien relacionada, aislada, etc.

F)Datos académicos

Este apartado hace referencia a los datos escolares que día a día irá aportando el alumno con su conducta escolar y que han de figurar necesariamente en su registro personal.

2. Constitución física y salud

La importancia de estos factores en el rendimiento y en la conducta escolar no es necesario exagerarla. Ciertamente que indicadores como desequilibrios talla-peso, variaciones en el crecimiento, problemas en los órganos sensoriales, anomalías en el habla, enfermedades, etc., son muchas veces factores causales de éxito o fracaso en el orden académico. De aquí la importancia de su exploración y detección.

A modo de resumen, el principal indicador con que debe contar cualquier ficha médicoescolar es el de crecimiento físico (longitud-peso, etc.).

Sabemos la significativa y estrecha relación que existe entre ambas variables. Por otro lado, las tipologías psicológicas y otras teorías muestran las correspondencias existentes entre variables de crecimiento físico y de comportamiento. De aquí que el estudio paralelo de crecimiento somático y desarrollo psicológico haya dado muchos y excelentes resultados y que la colaboración entre pediatras, pedagogos y psicólogos se haya demostrado no sólo vital, sino indispensable.

El campo somático objeto de observación es tan amplio y complejo que su estudio rebasa los límites de este trabajo, sin embargo, ello no implica que el maestro no deba ejercitar su "ojo clínico" para detectar posibles problemas vinculados a lo somático del individuo.

La siguiente cita de Brueckner y Bond (1961) - tomada de R.Sanmartín (1978) - refleja con claridad la importancia de vigilar estos aspectos: "Las condiciones físicas precarias, en general, la mala nutrición, enfermedades frecuentes y ciertos estados glandulares se relacionan, en cierto modo, con un rendimiento escolar pobre. Cualquier anomalía física del muchacho que desgaste su energía, le distraiga o le produzca inquietud o malestar, puede ejercer una perniciosa influencia sobre el aprendizaje. La atención difusa, la fatiga, el escaso control motriz, y los defectos visuales o auditivos tampoco contribuyen a la realización de un trabajo eficiente. Las características físicas y sensoriales del niño le predisponen a determinadas deficiencias en materias básicas como la lectura, escritura, cálculo, ortografía y otros aspectos del lenguaje" (pág. 61).

Como puede verse, un estado físico precario se relaciona con dificultades de aprendizaje en muchos sentidos. Sin embargo, no debe olvidarse que un niño sano, con dificultades escolares, presenta estos mismos síntomas, lo cual debe hacer dirigir la atención del profesor hacia el estado físico general. La salud del niño puede ser buena y, sin embargo, el cansancio, apatía o aburrimiento puede deberse a falta de energía física (poco descanso, mucha televisión, poco deporte, etc.)

3. Aptitudes o variables intelectuales

En el campo psicológico, se viene utilizando, según el momento histórico evolutivo y según distintos enfoques teóricos, distintos conceptos para hacer referencia a la posibilidad de realizar determinadas acciones que se califican de inteligentes: facultad, aptitud, potencia, factor, rasgo, capacidad.

Junto a una capacidad o dimensión intelectual general - inteligencia general - a la pedagogía le interesa más conocer aquellos otros factores intelectuales que permiten diferenciar a sujetos en función de diversos criterios tales como: contenido de la actividad, tipo de tarea, etc.

Es obvio que todo hombre, por el hecho de serlo, es inteligente en mayor o menor

magnitud. El problema histórico, sin embargo, ha sido el de ponerse de acuerdo en lo que entendemos por inteligencia. Muchas definiciones se han dado de tal variable de construcción (constructo), desde los griegos hasta nuestros días. Autores como Terman, Stern, Binet, Spearman, Thurstone, Cattell, Guilford, etc., han intentado definirla con más o menos acierto y según los instrumentos de medida que definieron para apreciarla.

Lo que es evidente es que al pedagogo le interesa conocer este mayor o menor grado de inteligencia que cada alumno tiene, porque esta facultad o variable está en la base de todo aprendizaje y rendimiento académico.

Como afirma R.Sanmartín (1978), desde 1890 en que Cattell utilizó la expresión de "medida mental', los trabajos experimentales en esta línea se han sucedido hasta que, a comienzos de siglo, Binet (1905) diseñó la primera escala de inteligencia. A partir de entonces se multiplicaron los trabajos y esfuerzos para medir la inteligencia, de tal modo que hoy en día contamos con numerosos tests para diagnosticar, con cierta fiabilidad, el grado de inteligencia de un sujeto.

Entre las unidades de medida más importantes, cuyo uso es frecuente, se encuentran:

- La edad mental. Unidad introducida por Binet, quien partió de la hipótesis de que a cada edad cronológica o real de un sujeto correspondía un cierto nivel de inteligencia. Para ello, definió cuestiones o pruebas características de una edad mental en función de que dichas cuestiones o ítems fueran resueltos por la mayoría de los sujetos de una determinada edad real o cronológica.

- Cociente intelectual: Unidad sugerida por Stern y que consistía en un simple cociente entre edad mental y edad cronológica:

$$CI = \frac{EM}{EC} \times 100$$

(La multiplicación por 100 se realiza para no operar con decimales).

- Puntuaciones centiles y tipificadas: Unidades técnicas construidas para interpretar puntuaciones brutas obtenidas en los tests con referencia a grupos amplios de sujetos (deciles, cuartiles, centiles, puntuaciones típicas, puntuaciones derivadas, etc.).

La aplicación de estos tests de inteligencia general permitió una primera clasificación de los sujetos en oligofrénicos, normales y superdotados, dependiendo de si su nivel intelectual está por debajo, a la par, o por encima de lo considerado normal en una determinada edad cronológica.

Esta diferenciación, pues, debe realizarse con tests y escalas de inteligencia (pruebas psicométricas), no obstante, cualquier profesor mínimamente experimentado sería capaz de situar a cada uno de sus alumnos en una u otra categoría, al menos provisionalmente, y si es preciso aconsejar el estudio de dicha dimensión intelectual por expertos.

Una vez aceptado un alumno como normal, es obvio que al educador le interesa más conocer, desde el punto de vista escolar y futura orientación académico-profesional, las aptitudes o variables intelectuales sobresalientes de cada escolar que el dato concreto del cociente intelectual. Por otra parte, el CI no es un valor fijo, sino que varía con la edad y el desarrollo madurativo del sujeto alcanzando su cúspide entre los dieciocho y veinticinco años según estudios realizados por autores como Otis, Weschler y otros. El carácter regresivo del cociente intelectual a partir de esta edad se debe a dos factores: la involución intelectual, por un lado, y la constante evolución de la edad cronológica, denominador del CI.

Con el desarrollo del análisis factorial a principios de siglo y con autores como Spearman, Thurstone, Cattell, Vernon, Burt, Guilford, Eysenk, etc., se descrubrió que junto a un factor general (g) o capacidad general del sujeto, la cual interviene en mayor o menor grado en cualquier clase de conducta inteligente, existen factores específicos responsables más directos de un rendimiento singular o específico. Son los denominados por Spearman factores específicos (s), más vinculados a los contenidos y tareas base de la medida del rendimiento intelectual. Sobre el número y características de dichos factores específicos o especiales existen diversas teorías fruto de la investigación empírica en psicología: Spearman ("g" y "s"); Thurstone (aisló mediante análisis factorial o factores intelectuales: comprensión verbal, fluidez verbal, razonamiento, factor especial, factor numérico y memoria, a los que denominó factores primarios); Cattell, Guilford, Eysenck, Vernon, entre otros, han definido distintas dimensiones intelectuales específicas más o menos coincidentes.

Lo que es claro es que al profesor le interesa conocer las aptitudes especiales de los alumnos, especialmente en orden a una orientación académica o profesional ulterior. Para ello y además de la observación diaria, deberá intentar explorar estas dimensiones enormemente relacionadas con la vocación del alumno y con su rendimiento escolar. Es evidente que la orientación de los escolares no puede basarse únicamente en las apreciaciones subjetivas del profesor. Estas técnicas (tests y otros instrumentos) deberán ser aplicadas o interpretadas por el profesor (tutor) convenientemente preparado y asesorado por especialistas en pedagogía o psicología.

La orientación escolar y profesional constituye un proceso que requiere de dicha exploración diagnóstica, que de modo global y comprensivo deberá realizarse por un equipo orientador constituido en el mismo centro por tutores y especialistas (médico, pedagogo, psicólogo y asistente social) en el caso ideal.

La vigencia de este servicio de orientación es cada día más plena en las disposiciones legales y su importancia se destaca especialmente a medida que se va acercando el final de la Enseñanza Primaria y en la Enseñanza Secundaria.

A título meramente indicativo el profesor puede consultar cualquier manual de tests o instrumentos de medida. En ellos se describen multitud de tests de fácil aplicación empleados en España para el diagnóstico de aptitudes intelectuales, intereses, personalidad, actitudes, etc.

4. Motivación, interés y actitud

El aprendizaje, como cualquier otro proceso cognoscitivo, tiene lugar como efecto directo de la actividad del sujeto que aprende. Ahora bien, dicha actividad será educativa, si es comprendida y querida por el sujeto. El alumno ha de ser movido interiormente a realizar la tarea académica mediante el impulso de la motivación. La motivación, los motivos, son elementos que según diversos autores despiertan, sostienen y dirigen la conducta. De aquí que la motivación se convierta desde esta perspectiva en factor central en la dirección eficiente del aprendizaje.

La función principal del profesor está en conseguir una actitud favorable, por parte del alumno, hacia los contenidos del aprendizaje, puesto que esta actitud condiciona positivamente esta adquisición.

Por otra parte, la fuente primaria y fundamental de la motivación es el interés, el cual viene a ser la especificación singular de la motivación. Como afirma R.Sanmartín (1978), el interés es el medio por el cual la mente es atraída hacia un objeto, o sea, el vínculo que une al sujeto con el objeto. Autores como Herbert, Claparede, Dewey, etc., han desarrollado un amplio cuerpo doctrinal sobre la importancia del interés o intereses como base de la pedagogía (pág. 36). Existen multitud de instrumentos de medida destinados a explorar y detectar intereses, entre ellos los más conocidos en España son los de Kuder, Strong, Thurstone, García Yagüe, etc., los cuales aparecen referenciados y definidos en los manuales de tests anteriormente mencionados.

Cercano al campo de la motivación y el interés está el terreno de las actitudes, cuyo influjo es decisivo en el rendimiento escolar. Ciertamente, sería preciso definir lo que entendemos por actitudes (muchas han sido las definiciones y operacionalizaciones de dichas variables). Sin entrar en una discusión teórica sobre ellas y tratando de sintetizar los matices de distintas formulaciones, podríamos decir que las actitudes son proyectos o predisposiciones para llevar a cabo una conducta determinada. En ellas influye sin duda la participación que el sujeto tiene en un determinado grupo social o cultural, de tal manera que puede decirse que son el resultado de aprendizajes realizados por la convivencia entre individuos. Las actitudes, pues, son disposiciones con una dirección y una magnitud; tienden a ser persistentes, pudiendo variar según niveles culturales,

sociales, económicos, profesionales y de ciertos grupos religiosos e incluso sociales. Referida a los escolares, la actitud o actitudes que un alumno puede adoptar frente a la constelación de estímulos educativos que la institución escolar ofrece puede variar en grado, desde una plena apertura, hasta una completa cerrazón o clausura. García Hoz (1970) afirma que "en muchas ocasiones, un bajo rendimiento se debe simplemente a una actitud negativa, que suele estar en relación, no sólo con la capacidad, sino con el interés que el estudiante tiene y también, en muchas ocasiones, con la relación personal que se ha establecido entre el estudiante y el profesor" (pág. 255). En la misma obra anterior, su autor delimitó una serie de campos dentro de los cuales es necesario observar las actitudes que el alumno muestra en su conducta. Estos campos, accesibles a la observación directa del educador, son:

a)Actitud frente al colegio.

b)Relaciones con los compañeros.

c)Relaciones personales.

d)Relaciones con las cosas.

e)Tono de vida.

Dentro de cada campo hay una serie de actitudes concretas agrupadas en apartados que son relacionados con sus actitudes contrarias o negativas (por ejemplo: originalidad-alteridad, delicadeza-grosería, alegría-tristeza, puntualidad-tardanza, confianza-desconfianza, cooperación-aislamiento, trabajo-pereza, constancia-inconstancia, etc.). En cada una de ellas, a su vez, pueden observarse distintos aspectos relacionados con: el hablar, el trabajo, el comer, el juego, los demás, las cosas, las normas, etcétera.

Lo que sí debe estar claro es que instrumentos de exploración como éste y otros similares no deben utilizarse con fines punitivos, sino, como afirman R.Sanmartín (1978) y García Hoz (1970), con fines encaminados a desarrollar al máximo las actitudes positivas, incidiendo de una forma más directa y personal sobre la propia clase y fundamentalmente sobre cada alumno a través de la tutoría.

5. Adaptación o integración social.- familiar, escolar y ambiental

No es necesario esforzarse demasiado en fundamentar la importancia del influjo del ambiente familiar y social en el desarrollo del niño y en su aprendizaje escolar. Es claro que a medida que el individuo se desarrolla tiene que encontrar nuevas formas de comportamiento adecuadas a cada fase de su evolución, de tal manera que se satisfagan sus necesidades y le permitan vivir equilibrada y armónicamente consigo mismo y con su entorno.

La vida misma supone un continuo proceso de adaptación al mundo, al entorno y a los problemas que necesariamente surgen en el desarrollo del individuo. Desde esta perspectiva, la adaptación supone una capacidad del individuo para orientar y percibir sus relaciones con situaciones, circunstancias, condicionamientos y la capacidad también de modificarlos. Determinados momentos en la vida del sujeto exigen la especial actuación de esta capacidad para adaptarse: al ingresar en la escuela y en las crisis propias de la evolución psicosomática de la pubertad y adolescencia, etc.

Suele decirse que una persona se halla adaptada, en la medida en que está en armonía consigo mismo y con su entorno. La adaptación no es mero conformismo, pero sí una aceptación de la realidad, aunque ello no implica necesariamente una actitud de cooperación con la realidad conocida. Puede hacerse uno cargo simplemente de su realidad o valorarla negativamente (e intentar modificarla); en ambos casos se acepta la realidad.

Dado que el hombre es un ser social por naturaleza, la adaptación abarca la capacidad de mantener relaciones con los demás que sean satisfactorias y duraderas. Estas relaciones implican el desarrollo de amistades y la participación en actividades de grupo, con el sentimiento de que uno es miembro aceptado de dicho grupo. Por el contrario, la inadaptación lleva al sujeto al aislamiento y a la clausura.

Tradicionalmente la noción de adaptación se refiere a lo social fundamentalmente. Desde el punto de vista psicopedagógico, corresponde al ajuste de los diferentes aspectos de la vida personal y social. Está vinculada en el niño de forma pasiva - como afirma R.Sanmartín (1978)a la capacidad de someterse al régimen dominancia-sumisión, pero lo propio de una correcta educación es desarrollar una adaptación activa, a fin de que el niño se realice, de acuerdo con sus potencialidades y el ideal ético y sociocultural de la familia y colectividad.

No vamos a entrar en un análisis más profundo del contexto histórico de la dimensión general de adaptación; simplemente señalaremos que existen distintos instrumentos de medida y valoración del grado de adaptación de un sujeto. El más conocido de todos es sin duda el construido por Bell para adolescentes y adultos, que ha sido traducido y adaptado a la población española por Cerdá (1978). Dicho instrumento permite obtener cuatro medidas distintas de adaptación personal y social (adaptación familiar, adaptación a la salud, adaptación social y adaptación emocional). Otros instrumentos que miden específicamente adaptación son los de Cerdá (adaptación escolar), Pérez Juste (adaptación social), Jiménez (adaptación escolar), etc. En general, todos los cuestionarios de personalidad son en buena medida cuestionarios de adaptación, en tanto que valoran integración personal, dimensión básica que implica desarrollo equilibrado y ajustada coordinación de capacidades, voluntad, emociones, etc.

Otros instrumentos también valiosos para conocer en qué medida un alumno se halla

integrado en la escuela y en su aula son las llamadas técnicas sociométricas. Dichas técnicas (tablas, sociogramas, psicogramas, etc.) permiten conocer la estructura interna del grupo: líderes, aceptados, rechazados, dependientes, independientes, etc. Su utilización es relativamente sencilla, aun cuando haya de cuidarse especialmente la interpretación de los resultados, pues si no se completan con otros datos procedentes de diversas fuentes (cuestionarios, entrevistas y observación directa) puede llegarse a conclusiones simplistas y superficiales que en vez de ayudar a entender la estructura de un grupo puede entorpecer su comprensión.

4.1.2. Evaluación continua

Si la evaluación inicial tenía por objetivo determinar el punto de partida de modo que sirva de referencia para la programación y la evaluación posterior, la evaluación continua (también llamada progresiva, diferida o sucesiva) es la verdadera evaluación del proceso educativo tomado en su conjunto. Su finalidad es determinar el grado en que se han conseguido los objetivos específicos de cada unidad didáctica o de aprendizaje en cada disciplina y en el conjunto del proceso educativo. No se trata de la evaluación basada en pruebas periódicas aplicadas, sino más bien la valoración permanente de la actividad educativa a medida que ésta se va desarrollando. Su carácter es, pues, dinámico al estar integrada en el propio proceso educativo. Siguiendo a R.Sanmartín (1978), la evaluación continua o progresiva persigue las siguientes finalidades (pág. 47):

a)Conocimiento de las posibilidades de los alumnos en orden a la valoración de su rendimiento y de la orientación escolar y profesional.

b)Conocimiento del proceso o desarrollo de la actividad educativa, valorando el conjunto de variables que inciden en ese desarrollo, planes, programas, métodos, técnicas, horarios, relación profesor-alumno, etc.

c)Ayuda constante al alumno, diagnosticando las deficiencias del aprendizaje y sus causas o por el contrario poniendo de relieve sus excelentes dotes (diagnóstico y pronóstico).

d)Conocer en qué medida se alcanzan los objetivos propuestos, el grado de progreso de los alumnos y la efectividad de la programación establecida.

e)Renovación crítica del sistema educativo, intentando conseguir la adecuada fijación de objetivos y la eficaz estructuración de los programas y actividades.

f)Información periódica a la familia de los resultados de la actividad educativa, estimulando la colaboración recíproca entre las dos entidades.

Como puede observarse, este tipo de evaluación descansa en una programación de

objetivos, contenidos y actividades que habrán de ordenarse y temporalizarse.

Esta formulación de objetivos debe tener dos niveles: mínimos y optativos. Los primeros implican un aprendizaje exigible a todos los alumnos de un nivel concreto, mientras que los segundos se refieren a aquellos contenidos elegidos libremente de acuerdo con la singular capacidad de cada uno.

Los programas se convierten así en guía del proceso educativo. Por ello no deben considerarse algo rígido y definitivo, sino flexibles, adaptables, dinámicos y abiertos a las características del nivel para el que han sido prefijados.

La evaluación, como actividad sistemática integrada en el proceso educativo, se realizará para el profesorado durante todo el período lectivo, sin interrumpir la marcha del trabajo escolar. Dicha evaluación se realizará basándose en:

a)Los trabajos y actividades realizados por el alumno en clase.

b)El resultado de la observación controlada sobre sus hábitos y actitudes.

c)El resultado de todo tipo de comprobaciones sobre los conocimientos, comprensión y aplicación de las distintas áreas.

Como vemos, por el mismo marco legal, la evaluación continua utilizará toda la información que pueda recogerse sobre el trabajo y rendimiento escolar del alumno. Es conveniente unir a ello una determinada organización del centro educativo que posibilite el acuerdo de criterios y modos educativos entre profesores de un mismo curso o grupo de alumnos. El mismo marco legal así lo sugiere al hablar del "equipo de evaluación" (equipo docente) y al establecer las sesiones de evaluación: "las sesiones de evaluación son las reuniones de trabajo de un equipo docente con el fin de estudiar y apreciar el aprovechamiento y la conducta de cada alumno del grupo, acordar las medidas de ayuda y recuperación que deban tomarse con los alumnos que lo necesiten, examinar la eficacia de los métodos adaptados y valorar los objetivos propuestos, así como el nivel de los contenidos".

El número de sesiones puede ser variable dependiendo del nivel de escolaridad o del país. Actualmente en Primaria y Bachillerato el marco legal habla de: "las sesiones de evaluación serán tres o cinco, regularmente espaciadas a lo largo del curso".

En todo caso, estas sesiones de evaluación no son más que cortes en el proceso evaluativo que permiten ir registrando resultados mediante pequeños periodos temporales. Ello no implica que inmediatamente antes de cada sesión se realice una prueba o examen. Esto mismo desvirtúa en cierta manera la filosofía de la evaluación progresiva o continua, que más bien trata de conocer en todo momento el desarrollo de la

actividad educativa, para de un modo rápido e inmediato corregir las deficiencias detectadas, sin necesidad de tener que esperar a la próxima sesión de evaluación. De este modo, las calificaciones en las sesiones de evaluación deben responder al aprovechamiento/rendimiento de los alumnos manifestado a lo largo del periodo que comprende y no a los resultados inmediatos y ocasionales de una prueba o examen (véase más adelante el concepto de evaluación formativa).

A diferencia del examen tradicional, esta evaluación no implica sólo al estudiante como sujeto, sino también y ante todo, al propio sistema escolar en su conjunto y a la pluralidad de agentes que interviene en toda acción educativa. La evaluación, por otro lado, no está circunscrita a un sólo acto, sino que se extiende a lo largo de todo un proceso y es sólo, en esa perspectiva, en la que adquieren sentido pruebas evaluadoras concretas. Al hablar de evaluación, por tanto, se hace referencia a procesos que permiten verificar la enseñanza y el aprendizaje al servicio de unos fines educativos precisos.

Es claro, pues, que la evaluación es un instrumento - de enorme importancia por su función de palanca del aprendizaje - al servicio de la educación, sin que pueda justificarse en ningún caso que la educación puede realizarse para la evaluación.

De todo lo anterior se deriva la necesidad de utilizar diversas técnicas e instrumentos para la recogida de información significativa que sirva de base para la evaluación progresiva. Fundamentalmente estas técnicas y procedimientos pueden clasificarse en dos: de observación (directa o indirecta) y de experimentación (todo tipo de pruebas que provocan la respuesta del sujeto ante estímulos definidos previamente). Esta clásica división de las técnicas de recogida de datos puede servimos perfectamente para encuadrar los distintos procedimientos de recogida de información que sirva de base para la evaluación continua. Las primeras (de observación) no modifican la situación natural de enseñanza-aprendizaje y están destinadas a apreciar ocasional o sistemáticamente conductas significativas que permitan conocer e interpretar el conjunto de variables que inciden en el sujeto y en el proceso educativo. Los segundos (de experimentación), por el contrario, tratan de provocar conductas, respuestas, implícitas en el proceso de enseñanza-aprendizaje. Ambos tipos de técnicas son de naturaleza muy distinta y, sin embargo, son complementarias y deben utilizarse conjuntamente para intentar apreciar con rigor la evolución del aprendizaje del alumno.

A modo de clasificación de las mismas, ofrecemos la figura 4.1. Las técnicas y procedimientos que en él aparecen serán desarrollados en un capítulo posterior dedicado exclusivamente a los instrumentos de medida y evaluación del rendimiento educativo y de otras variables cercanas al proceso educativo.

4.1.3. Evaluación final

En realidad, la evaluación final, como última fase del proceso evaluador, no es otra cosa

que la síntesis de todos los elementos proporcionados por la evaluación inicial y continua, con el objeto de llegar a la formulación de un juicio global que resuma el progreso realizado por el alumno durante el curso escolar. Es, pues, una valoración general, global y final del trabajo desarrollado por un alumno durante un periodo lectivo particular - un curso o un periodo mayor, una etapa, un ciclo, etc.-, y trata de ser una estimación cualitativa sintética referida a la cantidad/calidad de conocimientos y hábitos adquiridos al término de un periodo académico concreto.

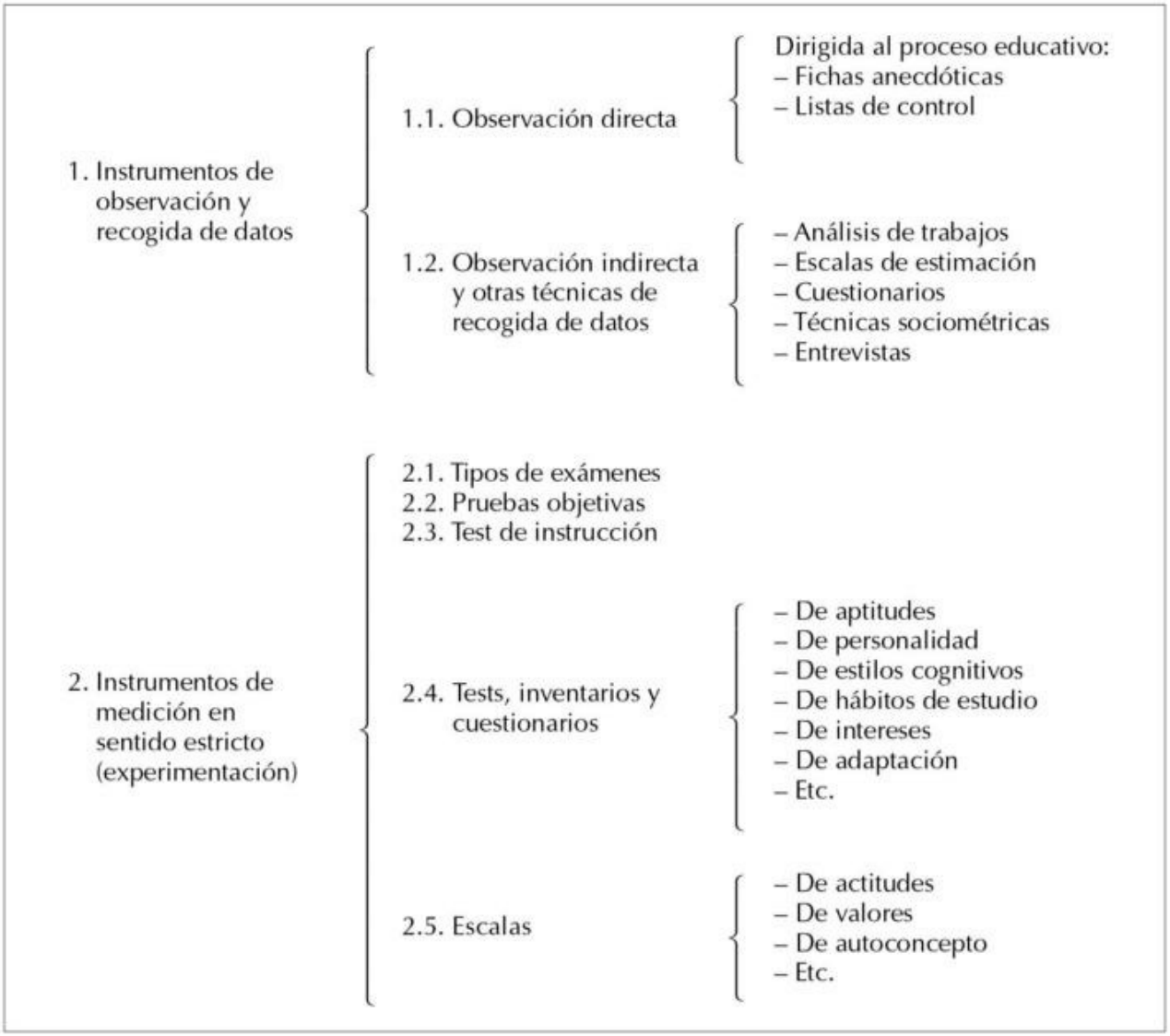


Figura 4.1. Instrumentos de recogida de información de cara al proceso evaluativo.

Si como hemos venido diciendo, la evaluación continua consiste en ir descubriendo y verificando, a lo largo del proceso educativo, las deficiencias y logros de un sujeto o de un programa educativo, para subsanarlas - las deficiencias - y evitar consecuentemente resultados finales negativos, es obvio que la evaluación final no puede ser otra cosa que

la resultante de todo este proceso, la consecuencia natural de la evaluación continua (debidamente informada a su vez por la evaluación inicial).

Es de destacar - como señala R.Sanmartín (1978) - que la evaluación final responderá a un juicio global de todos los profesores del alumno, porque es práctica bastante generalizada que cada profesor haga valer sus criterios respecto al área que desarrolla, perdiendo de vista la unidad funcional del alumno y la obligación de integrar el juicio de cada uno de los profesores intervinientes en una unidad colectiva superior (pág. 94). Por otra parte, no parece razonable desintegrar completamente a un alumno en sectores correspondientes a distintas áreas del saber. Sin embargo, habrá que cuidar también el extremo opuesto, tampoco razonable, de pensar que un alumno sobresaliente en un área o en varias ha de serlo en todas o siendo deficiente en una, o varias áreas, debe serlo en todas o en otras.

No obstante lo anterior y como afirma De la Orden (1982), "como factor supletorio y de contraste, es aconsejable aplicar al final de cada nivel una prueba flexible que, si está bien concebida, podrá añadir nuevas evidencias sobre la situación real del alumno; pero bien entendido, que la evaluación final es fundamentalmente el resultado ponderado de la evaluación continua" (pág. 32). En esta línea, la legislación vigente sobre evaluación habla, por ejemplo, de pruebas de promoción flexibles al final de cada uno de los cursos de la enseñanza primaria. Dichas pruebas, que serán preparadas, aplicadas y valoradas por un equipo de profesores designado por la dirección del centro respectivo, versarán exclusivamente sobre aspectos básicos y fundamentales del programa del nivel correspondiente. Si la valoración de la prueba coincide con los resultados de la evaluación progresiva, se otorgará al alumno la calificación correspondiente. Por el contrario, si existe discrepancia entre los resultados de las pruebas de promoción y la evaluación continua, se procederá a una entrevista con el alumno para decidir la calificación que proceda.

La valoración de las pruebas se realizará con una estimación global en términos de "sobresaliente", "notable", "bien", "suficiente" e "insuficiente". Con esta escala de cinco grados habrá una valoración ponderada en el supuesto de que sea positiva.

La inspección técnica de educación será la responsable de velar por que el nivel de contenidos, aplicación y valoración de las pruebas sea correcto, en aras a garantizar la calidad de la enseñanza.

La misma legislación mencionada en los párrafos anteriores continúa explicitando algunos criterios más respecto a la evaluación final y así afirma que, independientemente de los resultados obtenidos en cada uno de los niveles de enseñanza, la calificación final irá acompañada de un consejo orientador formulado por el equipo de profesores del alumno. Se realizará de acuerdo con las siguientes normas.

a)El consejo orientador se basará en los datos recogidos a lo largo de todos los cursos ya realizados por el alumno, y que deben figurar en su Registro Personal. Tendrá en cuenta en todo caso las aptitudes, los rasgos de su personalidad, su nivel de adaptación en la clase y, en general, todas sus circunstancias personales significativas.

b)El consejo recogerá, fundamentalmente, todos los aspectos de orientación personal y escolar importantes y, en especial, la conveniencia de promoción de curso, las actividades de recuperación y desarrollo que debe realizar el alumno y los aspectos que ofrecen dificultades a su progresión escolar.

El consejo de orientación emitido al terminar el último año de escolaridad contendrá, además, una estimación de las posibilidades futuras del alumno. Este consejo se formulará en todos los casos en términos positivos y, dado su carácter indicativo, se procurará ofrecer al alumno más de una alternativa.

Respecto a los alumnos de último curso de Enseñanza Primaria o de Secundaria obligatoria que no obtengan calificación final de "suficiente", la línea a seguir parece indicar que podrán optar por la repetición íntegra de dicho curso o por la realización de pruebas de madurez para obtener el título de Graduado Escolar o de Bachillerato. Dichas pruebas serán fijadas en su contenido específico por los centros, siguiendo éstos orientaciones emanadas del Ministerio de Educación y Ciencia y supervisadas por la Inspección de Educación, a la que deberán remitir un ejemplar de las pruebas aplicadas y un ejemplar de las actas de evaluación final.

4.2. Evaluaciones formativa y sumativa (funciones)

Desde el punto de vista práctico y aplicado conviene distinguir entre ambos términos (Bloom, 1969 y 1971, y Scriven, 1972 y 1973).

La evaluación formativa tiene como propósito mejorar un proceso educativo, un programa, en su fase de desarrollo. La evaluación sumativa se propone contrastar la eficacia general del proceso educativo ya desarrollado, esto es, la eficiencia de los resultados del aprendizaje o la eficiencia general de un programa ya desarrollado.

Como afirma De la Orden (1982), "las diferencias entre ambos tipos de evaluación responden a una doble consideración: el momento de su aplicación en el continuo del proceso educativo y las decisiones a que dan lugar, es decir, el uso que se hace de los resultados evaluativos en cada caso" (pág. 53).

En todo proceso global de evaluación resulta complejo, por lo general, separar con claridad la evaluación formativa de la sumativa. No obstante, también es cierto que la formativa - si existe - precede siempre a la sumativa. Esta última suele tener un carácter

final, es decir, se aplica al terminar un período lectivo concreto (trimestre, curso, etc.). El uso de la evaluación sumativa suele ordenarse a la toma de decisiones como las siguientes: selección de entrada a un curso, o a un programa o a una función; clasificación de los sujetos respecto a ulteriores alternativas a seguir (bachillerato, educación técnico-profesional, etc.); promoción/no promoción de un nivel educativo a otro; certificar competencias, etc.

Como vemos, este tipo de evaluación está íntimamente relacionada con las últimas fases del proceso de evaluación: la evaluación final y la orientación y toma de decisiones a partir de ella. En este sentido, Popham (1980) afirma que los evaluadores sumativos no deben implicarse, evitando formar parte del equipo que ha diseñado la enseñanza. En su sentido más dramático la evaluación sumativa sería el juicio final de una tarea educativa (pág. 24).

La evaluación formativa, por el contrario, está vinculada estrechamente a la evaluación continua o progresiva. De hecho, la evaluación formativa se aplica en los distintos estadios y momentos del proceso de enseñanza-aprendizaje y adopta formas muy diferentes. Puede consistir en una pequeña prueba escrita, en la observación del comportamiento discente, en el análisis de los trabajos de los alumnos, en una entrevista, etc. Su objetivo es proporcionar información a profesores y a alumnos y, a su vez, bases para la corrección de deficiencias encontradas en el proceso educativo. Puede también utilizarse para calificar, pero su función de ayuda al mejoramiento de la enseñanza y el aprendizaje será más efectiva, como afirma De la Orden (1982), si no se vinculan sus resultados a la toma de decisiones clasificatorias o promocionales. Actualmente diversos autores tienden a identificar evaluación formativa con evaluación continua, aun cuando el origen de ambos conceptos sea diferente.

Escudero (1980) afirma que existe cierta relación entre los conceptos de valor formativo y sumativo, resultantes de un proceso de evaluación, y los conceptos de validez interna y externa, respectivamente, con los que se juzga la calidad del proceso y de los resultados de un trabajo de investigación. Al evaluar formativamente un programa o proceso didáctico, se cuestiona su validez interna, si alcanza sus objetivos en la forma prevista, si tiene consistencia, si no falla como programa en sí. Al evaluar sumativamente, vamos un poco más lejos y cuestionamos de alguna forma su validez externa, su grado de generalidad, sus ventajas sobre otros programas, su eficacia no sólo para alcanzar unos objetivos, sino como programa en su conjunto en el que los objetivos previstos son, asimismo, cuestionados (pág. 29). En todo caso, conviene aclarar que ambos tipos de evaluación - en la idea Scriven (1969)- no se oponen en absoluto; de hecho, dicho autor no recomienda una por encima de otra, sino que afirma que ambas son necesarias.

Al hablar de usos de la evaluación y calidad de la enseñanza, De la Orden (1982) analiza fundamentalmente el valor de la evaluación formativa, vinculándola directamente

a la mejora de la calidad de los procesos de enseñanza-aprendizaje; de hecho, afirma que esta modalidad de evaluación "constituye la base de la enseñanza correctiva (recuperación) y la autocorrección de la acción del profesor y del proceso didáctico, cuando la situación es todavía reversible, mientras se desarrolla normalmente la secuencia enseñanza-aprendizaje y se puede canalizar con éxito hacia los objetivos previstos" (pág. 55). El mismo autor termina dicho capítulo afirmando que "la evaluación formativa, caracterizada por su directa relación con el proceso de enseñanza-aprendizaje, integrada en su propio desarrollo, puede tener un efecto altamente positivo sobre el aprendizaje de los alumnos, la acción docente del profesor, la organización de la clase, el uso del material didáctico, la orientación de los alumnos, la innovación educativa y muchas otras facetas de la enseñanza y de la educación. Su estrecha y directa relación con el propio desarrollo de la actividad educativa hace que el uso adecuado de sus resultados se convierta en el mejor instrumento para asegurar la coherencia entre objetivos, proceso y producto de la enseñanza y el aprendizaje. En otras palabras, los múltiples usos que caracterizan a la evaluación formativa, constituyen otras tantas vías para promover la calidad de la educación" (pág. 56).

Parece claro que ambos tipos de evaluación no se oponen, sino que se complementan, dadas sus características y objetivos diferenciales. En todo caso, el profesor debería dirigir sus mejores esfuerzos a la evaluación formativa; sin duda, su uso le dará enormes pistas y ayudas para mejorar la calidad de los procesos educativos en los que esté implicado. A través de ella, el evaluador formativo podrá recoger información y emitir un juicio sobre el valor de los diferentes aspectos de una secuencia educativa con el fin de mejorarla. Como afirma Pophan (1980), "el núcleo de la estrategia del evaluador formativo es recoger pruebas empíricas rela tivas a la eficacia de los distintos componentes de la secuencia pedagógica y analizar estos datos con el fin de aislar las deficiencias y sugerir modificaciones" (pág. 23).

Lo cierto es que hasta estos momentos la evaluación formativa ha sido poco utilizada en nuestros centros educativos de una forma sistemática e intencional.

4.3. Evaluaciones normativa, criterial y personalizante (sistemas de referencia en la evaluación)

Junto a la diferenciación anterior entre evaluación sumativa y formativa, otra distinción importante relacionada con ella es la que permite hablar de distintas instancias de referencia de la propia evaluación y así suele hablarse de evaluación normativa (comparación de la realización del alumno con la realización del grupo, muestra o población a que el alumno pertenece) y evaluación criterial (comparación entre la realización del alumno y la realización objetiva previamente establecida por vía racional o empírica). Es obvio que ofrecer un resultado de un instrumento de medida cualquiera sin un sistema de referencia es ofrecer una información inútil por muy precisa que sea. Los sistemas de referencia se utilizan, pues, para disponer de unas bases con las que

interpretar las puntuaciones de un instrumento de medida. Son la clave para que la información recogida sea útil.

El juicio de valor en que consiste toda evaluación básicamente implica la comparación de la realidad a evaluar con un patrón, instancia o criterio de referencia previamente establecido. Al hablar de evaluación del aprendizaje, y del alumno, la valoración de dicho aprendizaje necesariamente exige comparar las realizaciones del alumno con algún tipo de instancia de referencia. Y es aquí donde interviene la diferenciación entre las distintas instancias de referencia: bien la realización previa o las predicciones basadas en realizaciones previas (autorreferencia), o bien la realización determinada por factores externos al alumno (heterorreferencia) y dentro de ella referencia normativa o referencia criterial.

Hoy por hoy, el sistema o modalidad de heterorreferencia más utilizado es, sin duda, la referencia a normas. El profesorado considera las pruebas o exámenes que utilizan como base de la evaluación, como instrumentos de clasificación de sus alumnos intentando así ordenarlos según sus aptitudes, actitudes y rendimiento. Siguiendo a De la Orden (1982), el profesorado, al evaluar, muestra una clara tendencia a aprobar a los alumnos cuyas puntuaciones se sitúan en torno a la media del grupo, suspendiendo a los que no llegan a dicha media y asignando calificaciones de notable o sobresaliente a aquellos que obtienen puntuaciones superiores o muy superiores a dicha media del grupo. Otro dato que muestra también esta tendencia es el hecho de que cada profesor suele aprobar y suspender todos los años al mismo porcentaje aproximado de alumnos, lo cual no parece lógico dada la variabilidad conocida de los alumnos de curso a curso; esto es, el aprobado de un año no parece ser comparable, en cuanto a dominio objetivo de unos contenidos, con el aprobado del año anterior y del siguiente. En palabras de Pophan (1980), esto parece indicar que los profesores actúan como si la enseñanza no tuviera objetivos fijos y que el rendimiento de los alumnos, tal y como es sancionado por las notas finales, es un efecto de lo que los alumnos traen a la escuela, y se distribuye necesariamente de acuerdo a la curva normal. Como esto es lo esperado, el profesor no se siente fracasado ni responsable de aquellos alumnos que puntúen por debajo de la media. Es de justicia señalar, sin embargo, que muchos profesores intentan corregir esta perspectiva a través del establecimiento de mínimos que todos los alumnos deben conocer para aprobar una asignatura, independientemente de que la media o la mediana del grupo sea alta o baja. Es ésta una clara indicación - afirma Arturo de la Orden (1982)de una tendencia implícita en la educación hacia la evaluación criterial (pág. 73).

Otro tanto ocurre con los especialistas en medida educativa y constructores de tests empleados en la enseñanza; los tests estandarizados son ejemplos típicos de medidas referidas a normas. De hecho, siguiendo el modelo de la psicometría clásica - que se haya en la base de dichos tests-, el objetivo de los constructores de tests es intentar encontrar ítems o cuestiones que discriminen entre los sujetos, esto es, que diferencien bien. Estos tests tratan de encontrar diferencias entre los alumnos que son caracterizados

así por su posición relativa en el grupo (población normativa). Pophan (1980) afirma que el ideal es redactar ítems con un índice de dificultad de 0,50 (dificultad media) para así producir la varianza necesaria (diferencias máximas entre sujetos) al adecuado funcionamiento de los tests normativos. Además, las características técnicas de las pruebas - fiabilidad y validez - se determinan fundamentalmente por medio de coeficientes de correlación cuya eficacia como índices estadísticos exige una varianza relativamente amplia en las medidas y, por ende, en las variables que representan el rendimiento del alumno (téngase en cuenta que no es posible obtener correlaciones elevadas si las medidas son muy concentradas, esto es, si hay poca variabilidad). Al analizar esta situación, De la Orden (1982) afirma que: "una vez más, parece que los especialistas en medición educativa tienen una opinión singularmente pesimista respecto a la eficacia de la educación. Dan la impresión de no contemplar la posibilidad de que existan objetivos escolares que todos los alumnos pueden y deben alcanzar, siendo una consideración pedagógica irrelevante el hecho de que, en este supuesto, se reduzca considerablemente la variabilidad de las puntuaciones de las pruebas y, en consecuencia, su fiabilidad, validez y eficacia educativa. Si este fuera el caso, lo razonable sería abandonar este tipo de tests y no intentar mantenerlos distorsionando con ello el pensamiento pedagógico y la acción educativa" (pág. 74).

Puede decirse que la búsqueda de alternativas en los sistemas de referencia en la evaluación educativa se debe más a movimientos pedagógicos que a consideraciones psicométricas. Todos ellos coinciden en afirmar que la eficacia de la educación mejorará apoyándose en la idea de que con secuencias instructivas adecuadamente preparadas se puede lograr que la mayor parte de los alumnos alcance los objetivos de un curso o materia de enseñanza. "En este supuesto, las pruebas referidas a normas resultarían incongruentes para medir la efectividad de la instrucción, ya que si el tratamiento didáctico funciona, la dispersión de las puntuaciones en las pruebas se reduce, dificultando así interpretación con referencia al grupo" (De la Orden, 1982, pág. 74). En efecto, si se piensa que en la enseñanza se programan actividades dirigidas a promover cambios que culminen en el éxito del alumno, no debería extrañar que un profesor obtuviera resultados elevados en cuanto a éxito de los alumnos, por ejemplo, que un 95 por 100 o un 100 por 100 de los alumnos lograrán superar las metas previstas. Ello implicaría la no existencia de variabilidad o varianza en la variable de rendimiento, no pudiendo el profesor diferenciar a unos sujetos de otros, como pretendería una prueba referida a normas. Sin embargo, es evidente que la información conseguida es valiosa.

La evaluación de referencia criterial, como aportación, se debe fundamentalmente a Glaser (1962 y 1963), quien afirmaba que subyaciendo al concepto de medida del rendimiento educativo se halla la noción de un continuo de adquisición de conocimientos, que se extiende desde la absoluta incompetencia hasta la realización perfecta. El nivel de rendimiento de un alumno se define por su realización en la prueba, como algún punto de ese continuo. El grado en que el rendimiento del alumno se asemeja a la realización

deseable en un nivel especificado (objetivo) es determinado por medidas de aprendizaje referidas a un criterio. Pophan (1980 y 1983) define a una prueba de referencia criterial como "aquella que se emplea en averiguar la situación de un individuo con respecto a un campo de conducta bien definido" (pág. 149). De la Orden (1982), al precisar sobre el significado de la referencia criterial, afirma que, desde su aparición, la idea de tal modelo viene acompañada de una cierta confusión, cuya fuente principal es la multiplicidad de significados atribuidos a la palabra criterio, agravada por el hecho de que, en general, toda evaluación (juicio) implica referencia a un criterio. Pero justamente - continúa el mismo autor - es el carácter de este criterio lo que diferencia a este tipo de evaluación de la referida a normas (criterio normalizado en una población), dado que éste (el criterio) es su concepción prevalente, es caracterizado como campo o dominio de conducta bien definido (págs. 76-77). Obviamente, este tipo de evaluación requiere para su utilización la construcción de instrumentos de medida específicos, los cuales, en síntesis, exigen (De la Orden, 1982, págs. 79-80):

1. Definición de un campo o dominio de conducta claramente delimitado, que implica uno o varios objetivos de la enseñanza y el aprendizaje, formulados en términos de una clase homogénea de conducta (conocimientos, aptitudes intelectuales, etc.).

2. La selección de muestras representativas, dentro de cada clase, de conductas o realizaciones en forma de cuestiones o ítems que reflejan tales conductas.

3. La determinación y especificación de "estándares" o patrones de realización, como medida del criterio.

4. Establecimiento de un sistema de puntuación que describa adecuadamente la realización de los estudiantes en términos de aprendizaje (clase de conducta) bien definidos, independientemente de las realizaciones del grupo al que pertenece.

En todo caso, conviene aclarar que cada modalidad de evaluación - referida a normas o a criterios - está especialmente indicada para objetivos concretos, de tal forma que si se han de tomar decisiones clasificatorias o selectivas sobre sujetos, parece más aconsejable utilizar la referencia normativa, siendo la referencia criterial mucho más adecuada cuando se evalúa la eficacia de tratamientos didácticos, materiales o programas o bien el logro de unos objetivos mínimos de aprendizaje.

Por otra parte, y como señalan Beggs y Lewis (1979), "es lamentable que algunos educadores crean que deben elegir entre los dos sistemas de referencia cuando preparan una evaluación, porque una evaluación bien planeada debe emplear ambas formas de sistema de referencia. Un sistema de referencia es apropiado si está de acuerdo con los objetivos a evaluar" (pág. 25). De hecho, habrá objetivos que pueden requerir una comparación con un grupo conocido, mientras que otros objetivos pueden exigir la comparación con un nivel mínimo fijado por el evaluador. Ello indica que ambos

sistemas de referencia - a normas y a criterios - son apropiados y útiles en la evaluación educativa, pudiéndose utilizar independientemente o simultáneamente para obtener información útil en un proceso de evaluación. En general, el sistema de referencia ayuda al profesor a interpretar las puntuaciones recogidas. Ciertamente incumbe a profesores y evaluadores determinar un sistema para cada objetivo. ¿Cómo?; más adelante abordaremos la cuestión desde una perspectiva más práctica.

No son estas dos las únicas instancias de referencia posibles, sino más bien, como ya se dijo anteriormente, las dos modalidades de heterorreferencia o comparación con una realización determinada por factores externos al propio alumno. Ello no implica que la autorreferencia (comparación de la realización del alumno con la realización previa o con las predicciones basadas en realizaciones previas), no tenga interés; es más, desde el punto de vista de una concepción educativa personalizada, constituye sin duda uno de los elementos nucleares, al perseguir que cada alumno desarrolle sus potencialidades plenamente.

En esta perspectiva, se toma a la persona como "norma de sí mismo, de forma que el progreso consistirá justamente en el acercamiento a las metas particulares adecuadas a su situación de partida y a sus características. La evaluación en el marco personalizado debe llevar a que logros que merecen la valoración de satisfactorios para unos puedan ser claramente insatisfactorios para otros, incluso siendo estos objetivamente superiores a aquéllos" (García Hoz y Pérez Juste, 1984, pág. 67). En este contexto es en el que aparece la diferenciación entre los conceptos de rendimiento suficiente y rendimiento satisfactorio. El primero se refiere a la valoración de si el estudiante alcanza o no los niveles de aprendizaje previamente establecidos como mínimos (metas comunes) a todos los alumnos de un mismo nivel educativo; dichas metas comunes vienen dadas, de un modo mediato, por la política educativa y son concretados de forma próxima a la confección del programa por el profesor o equipo de profesores que han de desarrollar los contenidos en ese nivel. El rendimiento satisfactorio o insatisfactorio es el resultado de la valoración del rendimiento de un sujeto de acuerdo con su capacidad, con el fin de evitar los posibles desequilibrios entre sus resultados y sus posibilidades. Parece claro que si opinamos que dentro del período obligatorio de escolaridad es injusto exigirles a todos los alumnos unos niveles únicos, dado que implica medir a todos los alumnos por el mismo rasero, ignorando diferencias personales y ambientales, pensamos que es útil la consideración de la valoración del rendimiento como suficiente-insuficiente por un lado y como satisfactorio-insatisfactorio por otro. Obviamente, este segundo hace referencia indirecta a un modelo específico: la enseñanza individualizada o personalizada, la cual, sin olvidar la dimensión social de la persona, desciende hasta las características peculiares de cada individuo. Si queremos aprovechar los aspectos positivos que ambos rendimientos implican será necesario proponer un doble registro de resultados, según el cual, cada alumno es objeto de dos calificaciones. Dicho sistema, si se utiliza adecuadamente, ofrecerá tanto a la institución familiar como a la escolar, datos relevantes

sobre la evaluación del aprendizaje del alumno, pudiéndose adoptar medidas correctoras más acordes con la peculiar situación de cada alumno.

4.4. Otras modalidades de evaluación

Las clasificaciones anteriores de la evaluación son de las más importantes de cara a la práctica pedagógica. De hecho, la primera clasificación - inicial, continua y final - se identifica con el proceso mismo de la evaluación desde el punto de vista temporal. La segunda clasificación - formativa, sumativa - hace referencia a los objetivos específicos a los que se destina la misma evaluación, siendo la primera la más vinculada al proceso evaluativo entendido como acumulación de datos e información a profesores y alumnos y a su vez, bases para la corrección de deficiencias encontradas durante el proceso educativo. Se identifica prácticamente con la evaluación continua o progresiva, mientras que la evaluación sumativa suele ordenarse a la toma de decisiones al final de períodos lectivos concretos, esto es, se encuentra más vinculada a la evaluación final. La tercera clasificación, vinculada a las anteriores de alguna manera, distingue entre evaluación normativa, criterial y personalizada, siendo las dos primeras modalidades de heterorreferencia y la última, la evaluación con referencia a las características-posibilidades personales del sujeto (autorreferencia).

Estas clasificaciones anteriores no agotan, sin embargo, las modalidades de evaluación que pueden encontrarse en los manuales sobre la materia. Algunas otras dicotomías que suelen mencionarse son: heteroevaluación-autoevaluación, cuantitativa-cualitativa, orientadoraprescriptiva e interna-externa. La primera de ellas diferencia claramente entre la consideración del sujeto como objeto de la evaluación y la consideración del mismo como sujeto-objeto de la propia evaluación. Desde un punto de vista pedagógico es claro que interesa orientar al alumno hacia su propia evaluación. La autoevaluación fomenta la madurez personal y, a la vez, es signo de la misma. A través de ella, el que aprende va tomando nota de su propio aprendizaje, dado que es el principal interesado en él. Ciertamente, su incorporación a los procesos de enseñanza-aprendizaje exigen un cambio importante en la actitud didáctica que los profesores demostramos corrientemente, aunque tampoco parece conveniente tomar la autoevaluación como único criterio o modalidad evaluativa. Como modelo, la autoevaluación cuenta con defensores acérrimos y a la vez con numerosos detractores. En todo caso, cada profesor debe valorar en qué grado puede serle útil y rentable - pedagógicamente hablando - incluir la valoración del alumno de su propio trabajo, como un elemento más o como un elemento capital para implicar al propio alumno en la responsabilidad de su aprendizaje.

La dicotomía evaluación cuantitativa-cualitativa trata de distinguir entre modalidades cuyo énfasis fundamental se sitúa en la valoración de la "cantidad" de conocimientos, destrezas, hábitos, etc., desarrollados, o bien, en la valoración de la "calidad" de los mismos. Ciertamente, "los datos cuantitativos son más sencillos de medir y nos inclinaremos hacia ellos con más frecuencia. Resulta más costoso idear un plan de

valoración que detecte sutiles diferencias cualitativas. A pesar de todo, la importancia de estas variables más cualitativas justifican los esfuerzos del evaluador para medirlas con precisión" (Pophan, 1980, pág. 92).

Cabe también establecer una diferencia entre aquellas evaluaciones cuya meta es esencialmente orientadora (formativa) y otras más bien prescriptivas (de naturaleza sumativa fundamentalmente), dirigidas a proporcionar una conclusión precisa sobre determinados aspectos educativos. Finalmente, cabe distinguir entre evaluaciones realizadas por los mis mos agentes implicados en un proceso educativo (interna) y las que se efectúan desde fuera, a través de jueces externos al proceso educativo (externa).

En todo caso, coincidimos con la siguiente afirmación del documento elaborado por el Ministerio de Educación y Ciencia en su propuesta de Proyecto para la reforma de la enseñanza (1987): "el sistema educativo necesita de evaluación en todos esos ámbitos y modalidades. Es más, cualquier reforma de la ordenación educativa, así como cualquier propuesta pedagógicamente innovadora de renovación, tanto de la organización escolar como de los procesos educativos, ha de ir acompañada de un diseño de los procedimientos para su evaluación" (pág. 148). Es claro que dicha afirmación concede a la evaluación el peso que merece como motor y palanca de cualquier proceso de enseñanza-aprendizaje.

4.5. Otras funciones de la evaluación

De lo mencionado hasta aquí, puede observarse que la evaluación educativa abarca un amplio contexto de objetivos, en ningún modo reducibles a la mera constatación del nivel de rendimiento alcanzado por un alumno. De aquí que las funciones de la evaluación puedan ser muchas y diversas, incluso coincidentes con algunas de las modalidades anteriormente descritas. Por ejemplo, Scriven (1967), al diferenciar entre evaluación formativa y sumativa, lo que hace es distinguir dos funciones diferentes que debe cumplir la evaluación.

Por otra parte, la evaluación puede referirse al sistema educativo o a una parte de él, bien a un programa particular o en un centro, o bien a los procesos de enseñanza-aprendizaje de los alumnos (entre otros). La evaluación suministra información acerca del modo en que, de hecho, se desarrolla el currículum y se alcanzan los objetivos propuestos, información sin duda relevante para profesores, centros y administración educativa. En realidad, la justificación de la evaluación en el ámbito educativo radica en el hecho de que nos permite actuar sobre elementos de las estructuras y del proceso educativo, e incluso sobre los mismos objetivos previamente definidos, para modificarlos a la luz de los resultados. Es decir, "la evaluación nos permitirá conocer mejor los puntos fuertes y débiles de la enseñanza y del aprendizaje y, en consecuencia, facilitará su corrección y mejoramiento" (De la Orden, 1982, pág. 52).

La responsabilidad principal de los procesos de evaluación en los centros educativos debe correr a cargo, sin duda, de los profesores de los propios centros. "Los docentes han de evaluar críticamente su trabajo, analizar e identificar situaciones en las que el fracaso, la inadaptación o el bajo rendimiento del estudiante remiten a una disfuncionalidad del proceso educativo o a su propia inadecuación como profesor. El centro educativo ha de evaluar a los estudiantes sin perder de vista que un fracaso o inadaptación escolar generalizada es, ante todo, fracaso del propio centro escolar y de su proyecto educativo. Los servicios de inspección técnica deben desempeñar funciones que también son en rigor evaluadoras y orientadoras. Los profesores y los centros escolares han de utilizar los resultados de las evaluaciones, tanto los directamente obtenidos con sus estudiantes como los procedentes de evaluaciones externas, como un instrumento crítico que permite medir su propio trabajo docente y, cuando sea necesario, rectificarlo en la búsqueda de una práctica pedagógica mejor, más acertada" (Ministerio de Educación y Ciencia, 1987, pág. 149).

Así pues, la evaluación educativa en general, y la evaluación del rendimiento de los estudiantes en particular, debe cumplir algunas otras funciones importantes, entre las que suelen asignársele destacan las siguientes:

a) Función diagnóstica

Es propia fundamentalmente, aunque no exclusivamente, de la evaluación inicial. El diagnóstico de los distintos aspectos y elementos de la estructura, proceso y producto educativo, es la base de la adaptación de la enseñanza a las características y exigencias concretas de cada situación de aprendizaje y, a través de dicha función, podrá ser posible tomar las medidas correctivas y de recuperación precisas.

b) Función predictiva o de pronóstico

Sin duda alguna, el conocimiento de la situación de partida del estudiante, así como del rendimiento previo demostrado y de las posibilidades/aptitudes del alumno podrá servir de base para que el profesor intente predecir o pronosticar rendimientos y posibilidades futuras del mismo. Dichas estimaciones predictivas podrá realizarlas intuitiva o técnicamente (utilizando modelos estadísticos sencillos sobre ecuaciones de regresión simple o múltiple) pudiendo utilizarlas como base para su orientación personal, académica y profesional.

c) Función orientadora

La evaluación no puede conformarse con ser diagnóstica y predictora de limitaciones y ulteriores posibilidades y rendimientos. Si de hecho "descubre dichas limitaciones y posibilidades, permitirá tomar decisiones de ayuda o ayudar al alumno a que él mismo las tome. Programas correctivos específicos, ejercicios adecuados, clases de recuperación,

tutoría, son otros tantos elementos para ayudar a superarse, a la vez que se potencian las propias posibilidades" (García Hoz y Pérez Juste, 1984, pág. 69).

Esta función orientadora, como afirman los autores anteriores, probablemente se comprenda mejor por contraposición al tradicional valor selectivo de la evaluación, consistente en discriminar entre los alumnos que "valen"y que "no valen"; el carácter orientador parte de la base de que todos, en cuanto personas, valemos algo y para algo, y a la evaluación y a los evaluadores les corresponde ayudar a cada alumno a descubrir sus propias posibilidades, estimulando su pleno desarrollo.

d) Función de control

Esta función es inherente a la evaluación. A través de ella se asegura el control permanente del progreso educativo del alumno de cara a la información a la familia, a la realización ade cuada de las promociones de curso o nivel y, obviamente, a la expedición de títulos y diplomas sobre la base de datos lo más objetivos posible. La necesidad de conocer en qué medida se logra o no alcanzar los objetivos educativos previstos al programar los procesos de enseñanza-aprendizaje, implica la consideración de la función de control de la evaluación. Lógicamente esta función no tiene por qué hallarse descarnada de las otras funciones anteriormente mencionadas, es más, debe ir unida a ellas.

En resumen, y si el núcleo esencial de todo esfuerzo educativo, como es evidente, es el aprendizaje de los alumnos y su facilitación, las anteriores funciones de la evaluación - junto a otras posibles como "retroalimentación" del alumno, administración y supervisión, solución de problemas, etc. - apuntan directa o indirectamente a la facilitación, mejora y perfeccionamiento del aprendizaje (De la Orden, 1982, pág. 52).

5

La medida en la evaluación educativa

La medida, como venimos señalando, juega un papel básico en la evaluación educativa. Si la evaluación se basa en datos previamente recogidos, es obvio que esos datos habrán sido obtenidos por algún procedimiento que, en general, podemos denominar de medida. Sin medida, pues, no es posible hablar de evaluación o, al menos, de evaluación rigurosa y sistemática. Parece claro que si evaluar es, en último término, valorar o juzgar asignando una categoría a aquello que se evalúa, medir es justamente el procedimiento para definir, obtener y ofrecer información útil que posibilite el juicio o valoración posterior en que la evaluación consiste.

a) Cuestiones a responder

Ahora bien, cuando un profesor se enfrenta a la tarea de evaluar surgen distintas cuestiones a las que es preciso responder, previamente a la aplicación de medidas. Por ejemplo, ¿qué criterios voy a utilizar para asignar categorías o calificaciones?, ¿qué objetivos deben alcanzar los alumnos?, ¿qué nivel de dominio mínimo ha de alcanzar un alumno con respecto a uno o varios objetivos? La respuesta a estas cuestiones es capital y, de hecho, al responder a ellas se está definiendo la información que es preciso recoger para poder valorar el aprendizaje de los alumnos. Esto es, se está definiendo la información a obtener que será la base de la medida. Los instrumentos de observación o medida se encargarán de recogerla.

b) La información recogida debe ser útil

El profesor ahora deberá elaborar, aplicar y puntuar los tests o instrumentos de medida, proponer y puntuar al alumno en las variables base de evaluación, sean medidas a través de tests o a través de observación sistemática. Sin embargo, con esto no basta, la información ade más ha de presentarse de forma útil. Dicha utilidad dependerá no sólo de la bondad de los instrumentos utilizados (validez, fiabilidad, etc.), sino también de que la forma de presentación facilite su interpretación.

Los principios de la medida permiten proyectar estructuras evaluativas en las que puede incorporarse cualquier tipo de instrumentos y métodos de recogida de información: tests externos (no elaborados por el centro), pruebas internas (elaboradas en el centro), datos de observación, etc., de manera que las puntuaciones puedan ser interpretadas significativamente y ofrecidas en algún tipo de informe, "es decir, diseñados de modo que la información resultante de los exámenes sea significativa, comprensible y fiable" (Beggs y Lewis, 1979, pág. 20). Este es el objetivo de la medida en educación, proporcionar

información útil en el proceso de evaluación. Por otra parte y como ya se intuye, los principios de medición ofrecen las bases para juzgar la calidad de los instrumentos, proporcionan la posibilidad de interpretar las puntuaciones y los procedimientos para juzgar la calidad de los elementos o ítems de una prueba o test.

c) Los objetivos educativos

Para elaborar instrumentos específicos de medida que ayuden a obtener la información base de la evaluación, deben definirse previamente razones que justifiquen qué medir. Dichas razones no son otra cosa que los objetivos mismos de la educación, que de esta manera se convierten en objetivos de la evaluación. Son, pues, los objetivos los que determinarán el contenido y la forma que deberán tener los procedimientos de recogida de información.

d) ¿Cómo clasificar los objetivos de la educación?

La respuesta, conocida por todos, es: en generales, específicos y operativos. Los primeros, los objetivos generales, ofrecen muy poca ayuda a quien debe construir o decidir la elección del procedimiento adecuado de medida. Generalmente, no incluyen suficientes datos para determinar con qué criterios hay que compararlo, ahora bien, cumplen una función básica de orientación del aprendizaje que no debe subestimarse. De aquí que a partir del objetivo general sea preciso definir objetivos específicos. Un objetivo específico permite ya orientar el tipo de datos a recoger, concreta la información que será preciso obtener y ayuda a concentrarse en el problema de elaborar objetivos operativos o mensurables.

Veamos un ejemplo: supongamos que definimos un objetivo general como el siguiente: "avanzar en el aprendizaje del vocabulario". Este objetivo es obvio que no precisa demasiado el tipo de información a recoger. Se podría elaborar, sin embargo, y a partir de él, un objetivo específico como el siguiente: "el 90 por 100 de los alumnos de cuarto de Primaria aumentarán en diez puntos su puntuación directa en la prueba de vocabulario de García Hoz (1976) a lo largo del año de escolaridad". Este objetivo precisa mucho más el tipo de información a recoger, sin embargo, puede no ser satisfactorio por otras razones; por ejemplo, ¿es acertado el criterio de que sólo el 90 por 100 de los alumnos alcance este criterio mínimo?, ¿es tan espe cífico este objetivo que se ha diluido la intención del objetivo general? Como afirman Beggs y Lewis (1979), "cuando se redactan objetivos específicos, existe la tendencia a perder de vista la intención del objetivo general y a concentrarse sólo en el problema de la elaboración de objetivos mensurables. Si ocurre esto, los objetivos específicos no resultan significativos y las medidas o datos basados en ellos tampoco son significativos. Pueden fijarse objetivos específicos mensurables (operativos) cuando, de hecho, vayan a utilizarse instrumentos de medida, pero deben considerarse como subcategorías de los objetivos generales si tales medidas ofrecen información útil para las evaluaciones escolares" (pág. 24).

Los objetivos operativos, máximo nivel de concreción, permiten definir las conductas que esperamos se produzcan en el alumno una vez completado el proceso de aprendizaje. Son, pues, los que especifican la forma en que ha de medirse, los que describen los resultados que se pretenden obtener. Si, en general, un objetivo es un enunciado que describe un resultado de la enseñanza, más que un proceso o procedimiento de la misma, en el caso de los objetivos operativos ha de tener - siguiendo a Mager (1982) - tres características (pág. 23):

1. Realización de una acción (lo que el alumno debe estar en condiciones de hacer).

2. Condiciones importantes bajo las cuales deberá producirse la realización.

3. Criterio (calidad o nivel de realización que se considera aceptable).

e) Taxonomías de objetivos

Para la formulación de objetivos educacionales y también para la elaboración de elementos o ítems correspondientes a instrumentos de medida, una herramienta básica, sin duda, será tomar como base taxonomías de objetivos como las de Bloom y otros (1971), Gagné (1972), Kratworld y otros (1973), Guilford, etc., al objeto de explicar jerárquicamente los diversos niveles cognitivos o afectivos a conseguir. Quizás sea la taxonomía de objetivos cognoscitivos de Bloom y otros (1971) la más profusamente utilizada en nuestro país. En ella se especifican básicamente los siguientes niveles en el ámbito cognitivo (tomado de De Landsheere, 1977):

1. Adquisición de los conocimientos Evocación de hechos particulares y generales, de los hechos y de los procesos de una estructura o de un orden.

    1.1. Conocimiento de elementos particulares.

    1.1.1. Conocimiento de terminología.

    1.1.2. Conocimiento de hechos particulares.

    1.2. Conocimiento de los medios que permiten la utilización de los elementos particulares.

    1.2.1. Conocimiento de las convenciones.

    1.2.2. Conocimiento de las tendencias y de las secuencias, o de la evaluación de los fenómenos que se desarrollan en el tiempo.

    1.2.3. Conocimiento de las clasificaciones.

1.2.4. Conocimiento de los criterios, según los cuales son juzgados o comprobados los hechos, principios y opiniones, así como el comportamiento.

1.2.5. Conocimiento de los métodos.

2. Comprensión: Aquí se trata del nivel más elemental del entendimiento, que permite al estudiante el conocimiento de lo que es comunicado sin establecer necesariamente un vínculo entre dicho material y otros, o sin la captación de todo su alcance.

2.1. Transposición: Habilidad de transformar material o lenguaje técnico en términos más corrientes o en enunciados simbólicos y viceversa.

2.2. Interpretación: Explicación o resumen de una información. Se diferencia de 2.1 en que allí la transformación era literal, mientras que aquí no necesariamente lo es, incluso puede combinarse el orden o el enfoque.

2.3. Extrapolación: Extensión o generalización de un concepto, corriente o tendencia más allá del material presentado.

3. Aplicación: Utilización de representaciones abstractas en casos particulares y concretos. Estas representaciones pueden adquirir la forma de ideas generales, de reglas de procedimiento o de métodos ampliamente difundidos, o bien la de principios, ideas o teorías que convendrá recordar o aplicar.

4. Análisis: Implica separar o desmembrar los elementos o partes integrantes de una información, intentando aclarar la estructura jerárquica de ideas o las relaciones entre las ideas expresadas.

4.1. Investigación o análisis de los elementos.

4.2. Investigación o análisis de las relaciones.

4.3. Investigación o análisis de los principios organizativos.

5. Síntesis Implica la conjunción de elementos o partes con el fin de llegar a un todo. Esta operación cognitiva, inversa a la anterior de análisis, consiste en disponer y combinar los fragmentos, partes, elementos, etc., de manera que formen un plan o estructura que antes no pudo ser claramente distinguida.

5.1. Producción de una obra personal. Implica dar forma personal a una información, hecho o experiencia.

5.2. Elaboración de un plan de acción. Definición de una secuencia lógica (no memorística) de pasos o estructura jerárquica relativa a la resolución de un

problema o situación.

5.3. Derivación de un conjunto de relaciones abstractas.

6. Evaluación o valoración: Formulación de juicios de valor sobre el material y los métodos utilizados en un objetivo concreto. Incluye juicios cuantitativos o cualitativos sobre si el material y los métodos corresponden a los criterios.

Implica la utilización de una norma o criterio de apreciación que puede ser propuesto al estudiante o elegido/ establecido por éste.

6.1. Criterios internos. Valoración/evaluación de la exactitud/precisión de una información a partir de elementos críticos como son el rigor, la coherencia y otros criterios internos.

6.2. Criterios externos. Valoración/evaluación de la información a partir de criterios especificados o de otros que puedan recordarse.

Como toda taxonomía de objetivos, la clasificación anterior de objetivos educacionales es jerárquica, esto es, cada nivel superior implica la utilización o dominio de los niveles inferiores de objetivos.

En esta taxonomía, en la de Gagné (1972) o de cualquier otro autor, el profesor puede basarse para elaborar conductas formales que considere adecuadas para encauzar el aprendizaje de sus alumnos. Por otra parte, la misma clasificación de objetivos permite ponderar hasta qué punto los distintos niveles de objetivos que se persiguen en el desarrollo de los contenidos de una materia de aprendizaje se encuentran equilibrados unos con otros, evitando - si es deseable - que un nivel de objetivos predomine en detrimento de otros que se consideran también relevantes.

De cara a la evaluación de los alumnos y aunque las taxonomías no ofrecen comportamientos concretos directamente aplicables a la medida de alguna variable de rendimiento educativo, sin duda alguna pueden resultar de gran utilidad en la definición de objetivos operativos, ya que - como afirma Landsheere (1977) - las taxonomías "situadas a mitad de camino entre las grandes opciones ideológicas y los microobjetivos, parecen tender puentes entre la filosofía y la tecnología de la educación" (pág. 212). Para el redactor de elementos de un instrumento de medida del rendimiento, que prepara cuestiones pertenecientes a cada nivel cognitivo, de cada área de contenido y apropiados para cada nivel de experiencias previas, la taxonomía ofrece la base para construir un cuadro o tabla de especificaciones para cada grado o curso, lo cual le facilitará enormemente el diseño de un instrumento de medida equilibrado que pondere adecuadamente los diferentes niveles de objetivos a alcanzar para cada área de contenido que pretenda evaluar (cuadro 5.1).

CUADRO 5.1. Cuadro de especificaciones

NIVEL COGNITIVO	ÁREA DE CONTENIDO (LENGUA)			
	Vocabulario	Ortografía	Comprensión lectora	Expresión escrita
Conocimiento Comprensión				
Aplicación				
Análisis				
Síntesis				
Evaluación				

Para facilitar a los profesores la tarea de traducir en conductas concretas, evaluables y observables, las categorías o niveles definidos por Bloom, algunos autores han elaborado cuadros en los que junto a cada nivel taxonómico se ofrece una lista de verbos de acción y una lista de objetivos directos que, adecuadamente combinados, proporcionan el esqueleto de un objetivo operacional. Un ejemplo concreto - tomado de, De Landsheere (1977) y citado por Rotger Amengual (1984) es el siguiente, en este caso correspondiente a la categoría o nivel cognitivo de "aplicación" (figura 5.1).

Nivel	Verbos de acción	Objetivo directo
3. Aplicación	Aplicar, generalizar, enlazar, elegir, desarrollar, organizar, utilizar, emplear, transferir, reestructurar, clasificar	Principios, leyes, conclusiones, factores, ramificaciones, significaciones, corolarios, efectos, probabilidades

Ejemplo: Que el alumno sea capaz de transferir simbólicamente los efectos producidos por un volcán en erupción a la población y medio donde vive.

Figura 5.1. Ejemplo de correspondencia entre nivel cognitivo, verbos de acción y objetivo directo.

f) Limitaciones de la definición de objetivos

Conviene, en este momento, ofrecer también algunas reflexiones sobre las limitaciones de los objetivos de la educación. El hecho de que sean útiles no significa que no presenten algunos problemas y limitaciones que, conocidos, nos harán valorar en su justa medida sus posibilidades de cara ala medida y evaluación educativas.

Una primera limitación - como señalan Beggs y Lewis (1979) - es que desde el punto de vista del profesor, su desarrollo y cumplimiento es dificil y conlleva mucho tiempo. Puede, de hecho, parecer abrumadora para el profesor la tarea de sentarse en su mesa al principio del curso a delimitar y formular todos los objetivos a cumplir durante el mismo. Algunos profesores han señalado que los objetivos no deberían exigirse forzosamente, porque no tienen tiempo ni inclinación para formularlos (pág. 39). Sin embargo, hay que decir que sus ventajas parecen tan grandes que su elaboración no puede ignorarse. Quizás haya que motivar o motivarse en la tarea. Desde luego, una vez establecidos para un curso por un profesor o equipo de profesores, probablemente sólo será necesario revisarlos periódicamente.

Otro inconveniente, ya señalado, proviene de la necesidad de formular los objetivos específicos. Es fácil olvidarse del objetivo general al elaborar los específicos, lo cual puede llevar a desorientaciones y fragmentaciones de la unidad que supone el objetivo general. El profesor, como señalan Beggs y Lewis (1979), "debe recordar que los objetivos específicos son medios para alcanzar una meta más importante. Si se olvida esta meta y los objetivos específicos resultan fines en sí mismos, se pierden las ventajas de los objetivos educativos" (pág. 40). De la misma manera, los objetivos operativos no deben perder de vista el objetivo específico al que se ordenan.

También puede ser un problema serio, a la hora de elaborar objetivos educativos, la elección entre varios posibles. Ciertamente, pero es preciso decidir, aunque los fundamentos de la inclusión o exclusión de un objetivo sean, en ocasiones, poco claros.

"Debe hacerse una elección si se quiere que un programa educativo tenga objetivos sensatos" (Beggs y Lewis, 1979). Ahora bien, una vez elegidos los objetivos, los profesores deben evitar la tendencia a ignorar otros posibles resultados. Un ejemplo tomado de los autores mencionados más arriba puede ser muy clarificador. Supongamos que un profesor establece unos objetivos razonablemente alcanzables mediante instrucción programada o mediante discusiones en grupo. Una evaluación de ambas técnicas didácticas puede mostrar que las dos técnicas ayudan a cumplir los objetivos. El profesor, probablemente, elegirá la instrucción programada para cumplir esos objetivos en el futuro. Quizás esta decisión sea adecuada, pero también está olvidando que las discusiones en clase quizás han ayudado a los alumnos a alcanzar objetivos no intencionados: por ejemplo, a mejorar su aptitud para expresarse oralmente. Si ésta no fue una meta intentada en los objetivos, podría pasar inadvertida en la evaluación, pudiéndose perder una posible ocasión de desarrollar la aptitud de los alumnos para expresarse oralmente. Consecuentemente, parece aconsejable que los educadores estén

atentos, no sólo ante los resultados esperados, sino también ante aquellos que pueden aparecer sin esperarlos.

Finalmente, es preciso reconocer que, aunque los objetivos educativos pueden definir la conducta en términos muy precisos, no disponemos todavía de instrumentos precisos, fiables y válidos de medida de los diferentes aspectos de la conducta. De aquí se deriva la cautela de no "empeñarse" en disponer medidas objetivas muy precisas para todo tipo de objetivos educativos. De todas maneras, aunque no sean todos tan precisos, parece preferible formular objetivos que orientan los procesos de enseñanza-aprendizaje, que intentar abordar tales procesos sin objetivos. Probablemente, todo tipo de conductas o aprendizajes deseables pueden ser definibles, al menos, en términos de objetivos generales y específicos.

g) Elección del tipo de instrumento de medida apropiado

Llegado a este punto en el proceso de evaluación -ya han sido elaborados los objetivos-, el evaluador debe dirigir su atención a la forma en que va a recoger la información que precisa.

¿Cómo seleccionar la forma de recoger la información? La respuesta es obvia, depende de qué objetivo se trate. Habrá objetivos que probablemente podrán medirse utilizando tests estandarizados, esto es, instrumentos de medida externos al centro que se encuentran disponibles en casas especializadas en la construcción de instrumentos tipificados. Los más comunes son los tests de instrucción, aptitudes y actitudes. Aunque estén más lejos de la evaluación del rendimiento, entendido éste como producto individual inmediato de los procesos de enseñanza-aprendizaje, otros tests de personalidad, estilos cognitivos, intereses, etc., pueden ser útiles para recoger información de cara, fundamentalmente, a la evaluación inicial.

Habrá otros objetivos - probablemente la mayoría - que no podrán medirse a través de tests estandarizados, siendo preciso en tal caso, construir en el propio centro - el profesor o el equipo de profesores - instrumentos de medida ad hoc para medir dichos objetivos. A estas pruebas se las suele denominar instrumentos internos de medida; estos instrumentos, en esencia, no difieren de los externos, por ello habrá que construirlos siguiendo pautas similares. En los capítulos posteriores dedicaremos atención preferente a la construcción de estos instrumentos internos de medida, pues ello, pensamos, será de mayor utilidad para el docente en ejercicio. No obstante, también dedicaremos un apartado a la clasificación y valoración de los tests tipificados.

Dentro de cada uno de los grupos anteriores, podrían darse otras clasificaciones. Por ejemplo, dependiendo del sistema de referencia utilizado. Así, podemos hablar de instrumentos internos de medida de referencia normativa y de referencia criterial.

Finalmente, es preciso mencionar otros instrumentos de recogida de datos, relevantes para la evaluación educativa, no clasificables dentro de las categorías anteriormente mencionadas; nos referimos a instrumentos tales como:

-Instrumentos de observación sistemática.

-Ficha anecdótica y anecdotario.

-Listas de control y escalas de estimación.

-La entrevista.

-El cuestionario.

-Técnicas sociométricas, etc.

Sin duda alguna, no son técnicas para medir, en sentido estricto, el rendimiento educativo (entendido como producto individual inmediato de los procesos de enseñanza-aprendizaje), pero la información que pueden ofrecer de cara a la evaluación inicial, continua y final, es tan relevante que no debemos omitir una breve explicación de sus características, posibilidades y limitaciones.

Lo que parece claro es que los procedimientos para determinar la información apropiada a recoger son un paso clave en la evaluación educativa. Ahora bien, si los instrumentos de medida no son apropiados al objetivo específico o si el objetivo específico no refleja claramente el objetivo general, la evaluación carecerá de valor, aun cuando "técnicamente" los instrumentos estén muy bien construidos. Por otra parte, habrá que dedicar una atención especial a la interpretación de las puntuaciones de los propios instrumentos. Analizaremos este problema al desarrollar cada tipo concreto de instrumento de medida.

El procedimiento para construir y elaborar pruebas que permitan recoger la información necesaria y relevante para la evaluación educativa es complejo y lento. Pero no olvidemos que, si queremos dar la importancia que merece a la evaluación educativa como motor y palanca del aprendizaje del alumno, el tiempo y esfuerzo dedicados a determinar la forma de recoger la información no deberá escatimarse, pues de esta manera el proceso de medida estará perfectamente integrado en la evaluación educativa.

SEGUNDA PARTE

Técnicas e instrumentos de recogida de información y medida

Introducción

Las características y principios de la evaluación moderna dejan clara la imposibilidad de evaluar al alumno a través de una técnica o instrumento único. Si se desea evaluar el comportamiento global de un alumno, en sus distintos ámbitos (cognoscitivo, afectivo y psicomotor, por ejemplo) habrá que utilizar y combinar varios procedimientos.

Para la mayoría de los autores la distinción entre técnica e instrumento es importante. La técnica es el método de obtener informaciones, mientras que el instrumento es más bien el recurso concreto utilizado para obtenerlas. En este sentido, es más general el primer concepto, siendo el segundo una concreción o herramienta metodológica de la primera. De hecho, cuando se va a programar la recogida de información de base para la evaluación, habrá que seleccionar las técnicas a utilizar y los instrumentos que se usarán para cada técnica.

a)Sobre esta distinción podría hacerse una primera clasificación de los procedimientos y herramientas metodológicas útiles en la evaluación educativa. Las técnicas fundamentales serían básicamente tres: observación, encuesta y tests. Dentro de cada técnica los recursos instrumentales son, a su vez, múltiples.

b)Otros autores prefieren, sin embargo, distinguir entre tests y otros recursos metodológicos de recogida de información. Esto es, entre instrumentos de medida e instrumentos de recogida de información, basando la distinción en que los segundos no son instrumentos de medición en sentido estricto.

c)Una tercera clasificación distingue simplemente entre pruebas objetivas y pruebas de ensayo; esta clasificación se ocupa únicamente de la medida del aprovechamiento escolar, no incluyendo a aquellas otras variables que no miden rendimiento.

d)Otra clasificación útil es la que diferencia entre instrumentos externos e internos de medida, con clara referencia al hecho mismo del lugar en que la prueba se construye. Esto es, si el instrumento es construido dentro o fuera del centro en que va a ser aplicado.

e)Una clasificación más, diferencia dentro de las pruebas de medida del rendimiento, según el tipo de referencia utilizado y así puede hablarse de instrumentos de referencia normativa y de referencia criterial.

Algunas clasificaciones más podrían mencionarse. En esta obra vamos a utilizar una clasificación basada en varias de las anteriores mencionadas, dedicando en la explicación posterior especial atención a la medida del rendimiento a través de pruebas internas o externas, de referencia normativa o criterial y un menor énfasis en las técnicas de

observación, encuesta y en los tests tipificados, cuyo objetivo directo no es la medida del rendimiento o aprovechamiento escolar.

Seguramente la clasificación anterior no es exhaustiva, ni tampoco comprensiva de todos los posibles instrumentos que para la recogida de información relevante para la evaluación educativa podrían enumerarse. Sin embargo, sirva como una primera aproximación que iremos desmenuzando en posteriores apartados.

La primera división que aparece en el cuadro de la página siguiente es la que permite distinguir entre instrumentos de recogida de datos y de medida. Probablemente no todos los autores estén de acuerdo con la misma, pues sus límites son imprecisos. No obstante, hemos querido realizarla para diferenciar entre instrumentos que responden claramente al concepto de medida (acto de medición) e instrumentos que no responden estrictamente a tal concepto. En un capítulo introductorio definimos - siguiendo a Stevens - el acto de medición como el acto por el cual se atribuían números (numerales) a los objetos o relaciones empíricas de acuerdo con ciertas reglas. Para poder atribuir tales "numerales" o símbolos, es preciso disponer de instrumentos o técnicas en las que se puedan registrar manifestaciones de lo que se quiere medir. En este sentido, "los instrumentos o técnicas pueden ser ellos mismos el elemento material de la medición o tratarse simplemente de materializaciones en las que el investigador o los sujetos registran la existencia o no y el grado en que se da el hecho o relación que se investiga" (Pérez Juste, 1983, pág. 79). Efectivamente, cuando construimos o utilizamos un test de rendimiento o aptitudes, por ejemplo, las cuestiones o ítems del mismo, en su sentido más amplio, ponen al sujeto en situación de evidenciar la posesión de determinados conocimientos, habilidades, destrezas y nivel logrados; con ello se está midiendo realmente la variable tal como es apreciada por la prueba (rendimiento o aptitud). Sin embargo, en otras ocasiones nos encontramos en realidad ante algún tipo de material que sirve para registrar si una conducta se da o no y en qué grado en los sujetos. Ello ocurre, por ejemplo, en las listas de control, en el anecdotario o en la entrevista, aunque con claros matices diferenciales. En estos casos, más que medir, se están registrando datos útiles para una posterior elaboración o interpretación.

- A) Instrumentos de recogida de datos
  - Observación (Instr. de registro)
    - Observación directa
      - Anecdotario y ficha anecdótica
      - Listas de control
      - Escalas de estimación
    - Observación indirecta
      - Análisis de trabajos
      - Escalas de producción
  - Encuesta
    - Cuestionarios e inventarios
    - Entrevistas
    - Técnicas sociométricas
- B) Instrumentos de medición en sentido estricto
  - De aptitudes (tipificados, externos)
    - Inteligencia general
      - Verbal
      - No verbal
      - Mixtos
    - Aptitudes diferenciadas
      - Baterías
      - Razonamiento
      - Creatividad/Imaginación
      - Aptitudes perceptivas y de atención
      - Memoria
      - Destreza manual y motora
      - Aptitudes profesionales
      - Aptitudes espaciales, numéricas, verbales y artísticas
  - De personalidad, intereses, actitudes y adaptación (tipificados, externos)
    - Cuestionarios e inventarios de personalidad
    - Escalas de actitudes
    - Tests proyectivos y clínicos de personalidad
    - Escalas de actitudes
    - Pruebas de adaptación: escolar, personal, familiar y social
    - Intereses vocacionales y profesionales
    - Estilos cognitivos y otras variables
  - De rendimiento y pedagógicos
    - Externos (tipificados)
      - Baterías de pruebas pedagógicas
      - Pruebas de lectura, escritura, composición, ortografía, dibujo, etc.
      - Técnicas y hábitos de estudio
    - Internos (construidos por el profesor o centro)
      - De referencia normativa (pruebas objetivas o de ensayo)
      - De referencia criterial (pruebas objetivas o de ensayo)

Clasificación de las principales técnicas e instrumentos de recogida de datos y medida.

6

Instrumentos de registro de datos

Siguiendo la clasificación expuesta en la página 91, dedicaremos este capítulo a enumerar y explicar brevemente los principales instrumentos que pueden utilizarse para la recogida de información mediante el registro de datos.

6.1. Registro de datos de observación

Observar es mirar, oír, intentando prestar atención selectiva hacia determinados aspectos importantes y relevantes de una situación o de la conducta o de un sujeto o grupo.

La observación es, sin duda alguna, una de las técnicas que pueden permitir al docente conseguir informaciones directas acerca de las habilidades cognitivas, afectivas y psicomotoras de sus alumnos. Pero está especialmente indicada en estos dos últimos casos. Sus ventajas respecto a los instrumentos de medida (tests) son claras, especialmente en el hecho de no modificar la situación natural y permitir una interpretación global y comprensiva de una conducta o situación. Domingues (1977) sintetiza en los siguientes elementos o variables la utilidad e idoneidad de la observación al intentar evaluar (pág. 44):

-Actitudes, hábitos de estudio, adaptación social, cualidades de liderazgo, etc.

-Habilidades físicas como nadar, saltar, correr, etc.

-Habilidad de lectura oral: oír, realizar un experimento, bailar, tocar un instrumento musical, etc.

-Comportamientos típicos de los alumnos al hacer un examen, participar en una discusión, estudiar en grupo, etc.

Así pues, la observación intenta apreciar atentamente los fenómenos objeto de estudio sin intentar modificarlos. La observación del alumno está en la base de todas las actividades educativas y didácticas del profesor, pedagogo o educador. Como afirma R.Sanmartín (1978), "constituye el fundamento de la tarea orientadora del profesor y la condición previa de una enseñanza basada en métodos individualizantes" (pág. 52).

Sería, a mi juicio, un error que el profesor basara todos sus juicios y orientaciones a los alumnos en datos provenientes únicamente de mediciones de variables pedagógicas o

psicológicas. Se precisa de una labor sistemática de observación, que exige del profesor el dominio de los principios y categorías del ver y del comprender, que conozca los métodos de diagnóstico, medios auxiliares y prácticas psicopedagógicas a través de su formación pedagógica (en modo alguno significa esto infraestimar el valor de los instrumentos de medida, sean estos externos o internos, tipificados o construidos por el profesor, de referencia normativa o criterial).

Al intentar clasificar las técnicas e instrumentos de observación de cara a la evaluación educativa (fundamentalmente en su fase progresiva o de evaluación continua), suele distinguirse entre:

a) Observación directa (dirigida al proceso educativo)
   - Anecdotario y ficha anecdótica.
   - Lista de control.
   - Escalas de estimación.

b) Observación indirecta (análisis de tareas y trabajos escolares)
   - Análisis de tareas y trabajos.
   - Escalas de producción escolar.

6.1.1. Observación directa

Dirigidas al proceso educativo en su conjunto, trata de estudiar al alumno como sujeto activo de su propia formación dentro del aula.

Para que la observación directa constituya una herramienta útil y valiosa de recogida de información, ha de ser realizada con unos objetivos claros y definidos, ha de ser convenientemente planificada y desarrollada sistemáticamente. Debe intentarse registrar datos con posibilidades de efectuar comprobaciones posteriores.

a)Momentos Cualquiera de las situaciones de la vida escolar: conversaciones en clase, discusiones en grupo, trabajo personal en clase, trabajo de grupos, deporte, recreo, excursiones, etc.

b)Conductas o variables: Aquellas que se consideren relevantes desde el punto de vista de la formación integral del alumno. Podrían incluirse: conductas físicas, higiene, conducta en el trabajo, actitud personal hacia el estudio, hacia compañeros y profesor, asistencia, puntualidad, relaciones intraescolares y cuantas conductas o variables considere relevantes el profesor. Obviamente, puede partirse para la observación de algún problema detectado: falta de rendimiento esperado, deficiencias particulares del individuo o problemas académicos.

Desde luego, un primer paso podría ser el interés del profesor por conocer la historia personal de cada niño; tanto la historia pasada (a través de entrevistas con los

padres o a través de conversaciones naturales de la vida escolar), como la presente (situación social, cultural y educativa del contexto familiar). Puede ser muy útil, en este sentido, la información que el profesor puede recabar del orientador o pedagogo del centro o del asistente social.

c)Planificación de la observación: Como hemos expuesto, el campo de observación es para el profesor inabarcable. De ahí que sea preciso, para que la observación constituya una herramienta útil, que sea planificada, programada y desarrollada sistemáticamente.

La primera cuestión que se le plantea al observador, como afirma Sanmartín (1978), es la determinación del campo de actuación y dentro de éste seleccionar con precisión:

-Qué es lo que se va a observar.

-Qué aspectos se van a observar detalladamente.

-Qué o cuáles sólo de un modo general o global.

Será, pues, necesario confeccionar una lista de aspectos particulares del fenómeno, conductas o tareas que se están desarrollando y que nos interesa observar. Para ello, es necesario establecer unidades de observación de acuerdo con las siguientes normas:

-Formulación clara de las unidades.

-Adecuación al objeto, es decir, implicación de aquello que se quiere observar, validez.

-Que cada una de estas unidades represente un comportamiento más o menos frecuente y duración de los hechos observados.

-Que representen los aspectos más significativos de una situación concreta, ¿cuáles son los aspectos más importantes que se deben observar?

Un problema que todo observador se plantea, llegado este momento, es el de ¿cómo y cuándo registrar estos datos? Es evidente que el registro de los hechos, mientras están ocurriendo, es el más fiable al evitar problemas de memoria y la interpretación del observador, pero no siempre es fácil. Asti Vera (1972) recomienda, por ejemplo, "hacer anotaciones breves en una hojita de papel, disimulada en la palma de la mano o en el forro de un paquete. Si el registro exige anotaciones más extensas, el observador puede retirarse brevemente". Para contrastar los datos recogidos, cuando ha tenido que realizarse de forma diferida (no en el momento de

producirse el hecho), un procedimiento adecuado puede ser la observación intersubjetiva simultánea (realizada por varios observadores), lo que permitirá comparaciones posteriores de los datos recogidos.

d)Instrumentos Entre las herramientas habituales de observación directa suelen mencionarse fundamentalmente: anecdotario, listas de control y escalas de estimación.

1. Anecdotario y ficha anecdótica.

La ficha es, simplemente, una cartulina en la que se van anotando acontecimientos, incidentes significativos y relevantes con relación a un alumno determinado.

Lo más importante del registro de una anécdota es separar el hecho de la opinión o interpretación del mismo. El observador debe anotar el hecho - sin opinar sobre él- en el anverso de la ficha. Si se desea interpretar, lo hará en el reverso de la misma. Junto al hecho deben aparecer otros datos de contextualización, tales como el día, la hora, el lugar y cuantas circunstancias se consideren relevantes.

El anecdotario no es más que un archivo o fichero de anécdotas sobre un mismo sujeto o bien sobre una conducta particular del sujeto, que está siendo objeto de observación.

El valor de ambas técnicas está en la riqueza de la información que ofrecen. Dicha información está contextualizada y matizada por las mismas circunstancias que constan en cada anécdota. Su valor desde el punto de vista longitudinal puede hacerla especialmente útil para seguir la evolución de una conducta particular o bien de un problema específico de un alumno.

ANVERSO

Alumno Curso

Fecha Hora

Descripción del hecho

REVERSO

Comentario

Observador

Figura 6.1. Modelo de ficha anecdótica.

2. Lista de control

Está definida por la enumeración de una serie de conductas cuya presencia o ausencia en un grupo de alumnos se desea constatar. El observador sólo debe marcar con una cruz (x) si el hecho se da o no se da durante un período de observación definido. No hay comentario alguno sobre la conducta, ni apreciación del grado en que se da o no se da. En estas características radican sus ventajas y sus limitaciones.

Un modelo de listas de control - tomado de I.C.E.U.M. (1976) y citado por R.Sanmartín (1978, págs. 59-60) es el siguiente (figura 6.2):

Figura 6.2. Modelo de listas de control.

1. Conducta del alumno antes y durante la realización de sus trabajos.

Alumno
Materia o tarea
Fecha de este registro

	Sí	No	Dudoso
Receptividad:			
– Se hace cargo			
– Actitud de apertura			
– Atención selectiva			
Reactividad:			
– Acepta "las reglas del juego"			
– Acepta la responsabilidad:			
• Ante las tareas obligatorias			
• Ante las tareas optativas			
– Alegría ante el trabajo			
Realización:			
– Estabilidad emocional			
– Coherencia			
– Interés en la tarea:			
• Inicial			
• Sucesivo			
– Buenas relaciones con el grupo			
– Tipos de actividad que realiza:			
• Actividad de expresión verbal			
• Actividad de expresión matemática			
• Actividad de expresión plástica			
• Actividad de expresión dinámica			

2. Chequeo de los hábitos de estudio y trabajo intelectual

Alumno
Materia o tarea
Fecha de este registro

	Sí	No	Dudoso
Cognoscitivos:			
– Conoce el vocabulario específico			
– Aprecia los procesos y tendencias			
– Usa clasificaciones y categorías			
– Utiliza las ideas adquiridas			
– Es capaz de traducir a otras formas de lenguaje el material que se le facilita			
– Sus conclusiones están de acuerdo con los datos			
– Organiza y reorganiza el material recibido			
– Aplica las teorías y métodos generales a situaciones particulares y concretas			
Operaciones:			
– Distingue lo principal de lo accesorio o secundario			
– Distingue los hechos de las hipótesis u opiniones			
– Aprecia la unidad de la comunicación			
– Sus generalizaciones son lógicas			
– Afirma con documentación y/o pruebas			
– Afirma atendiendo a teorías previas			
– Usa bien los distintos tipos de lenguaje			
Trabajo intelectual:			
– Lee			
– Estudia			
– Escucha			
– Observa			
– Hace anotaciones y las observa			
– Se expresa oralmente			
– Resume y/o amplía			
– Recoge datos e informaciones			
– Usa bien el material auxiliar			
– Utiliza libros de referencia o consulta			
Habituales:			
– Comprueba su propio progreso			
– Exactitud y orden en el trabajo			
– Organiza bien su tiempo			
– Ejecuta su trabajo formalmente bien			
– No necesita que se le recuerde su deber			
– Vence por sí mismo los obstáculos que encuentra			
– Sabe trabajar en "gran grupo"			
– Sabe trabajar en "equipo"			
– Se responsabiliza de sus trabajos y opciones			
– Ayuda a los demás			
– Coherencia entre pensamiento y acción			

3. Escalas de estimación

Como afirma Pérez Juste (1983), "las escalas de estimación vienen a suponer, de alguna manera, la unión de los aspectos más positivos de la ficha anecdótica y de las listas de control, eliminando o reduciendo sus puntos más débiles". En síntesis, son listas de control cualificadas, esto es, una enumeración de conductas a observar seguida de una graduación o contínuum que describe la manera, forma o grado en

que se manifiesta cada característica a observar en cada alumno. Esta técnica se viene utilizando también en alguna de sus variantes (numérica, gráfica o verbal), para la estimación y medida de variables actitudinales, de intereses, de aficiones, valores, etc.

La gradación puede estar formada por tres o más niveles o valores que suelen abarcar la totalidad del continuo o amplitud de la variable o conducta que se estima. Modelos de ítems correspondientes a las distintas clases de escalas de estimación pueden ser los siguientes:

1. Escala numérica (el valor más alto indica máximo nivel):

– Sociabilidad 1 - 2 - 3 - 4 - [5]

– Agresividad [1] - 2 - 3 - 4

2. Escala gráfica (mayor número de subdivisiones):

– Sociable

											×
1			2			3			4		5

– Agresivo

×											
1			2			3			4		5

3. Escala verbal (expresión verbal de los grados):

– Sociable

×				
Mucho	Bastante	Normal	Poco	Nada

– Agresivo

		×	
Siempre	Bastantes veces	Pocas veces	Nunca

– Sociable

		×	
Mucho	Bastante	Poco	Nada

En estas escalas, los adjetivos que expresan los distintos grados pueden apreciar descriptivamente.

-Simplemente grado.

-Persistencia.

-Frecuencia.

-Intensidad.

-Etc.

Las conductas pueden explicarse en la propia escala para facilitar la asignación del grado.

A modo de ejemplo de distintas escalas descriptivas utilizables en un centro para apreciar conductas o variables relevantes para la evaluación, puede observarse en el cuadro 6.1.

6.1.2. Observación indirecta

Con la observación indirecta se hace referencia al estudio y análisis del producto de la actividad escolar, considerado en sí mismo. De aquí que suelan mencionarse dos técnicas: el análisis de tareas o de trabajos (incluye el análisis formal o informal del contenido de los trabajos) y las escalas de producción escolar.

a)El análisis de tareas exige una preparación previa minuciosa que fije con precisión los aspectos relevantes que van a ser considerados en la evaluación de una tarea o trabajo. Estos aspectos, en el fondo, son criterios de valoración de la calidad de un trabajo: contenido, ordenación de ideas, forma de expresarse, riqueza de vocabulario, estructura gramatical de las frases, corrección ortográfica, presentación del trabajo, ilustraciones y ejemplos, brevedad, claridad, juicio crítico, documentación del trabajo (bibliografía), estructuración propia, etc.

Es evidente que el análisis formal de un contenido debe realizarse siguiendo normas muy definidas. Dos ejemplos de aplicación que señala R.Sanmartín (1978), tomados de I.C.E.U.M. (1976), y García Hoz (1973), respectivamente, pueden ilustrar los distintos aspectos o elementos a valorar (obviamente, dichos aspectos variarán según el contenido de la materia y el criterio de valoración del profesor) (cuadros 6.2 y 6.3).

CUADRO 6.1. *Escalas descriptivas para apreciar conductas (Pérez Juste, 1983, pág. 92)*

ELEMENTOS DE OBSERVACIÓN	GRADOS DE VALORACIÓN INTUITIVA					MESES									TOTAL
	1	2	3	4	5	Oct.	Nov.	Dic.	Ene.	Feb.	Mar.	Abr.	May.	Jun.	
1. ASPECTO EXTERNO	Muy descuidado	Algo descuidado	Porte normal	Muy presentable y distinguido	Presencia fina										
2. EXPRESIÓN ORAL	Indistinta, fastidiosa	Desagradable, destemplado	Clara y agradable	Precisa, armónica	Correcta, fácil y rica										
3. SOCIALIBIDAD SOLITARIO	Aislado del grupo	Pasivo en compañero	Bueno y servicial	Amigo alegre	Jefe activo										
4. DISCIPLINA	Rebelde, agresivo	Alborotador, no dócil	Sumiso	Obediente	Colabora activamente										
5. CAPACIDAD INTELECTUAL	Escasa y lenta	Apenas suficiente	Normal	Buena	Muy buena										
6. APLICACIÓN AL ESTUDIO	Inactivo y perezoso	Inconstante, con fatiga	Normal	Bueno y con interés	Máxima y constane										
7. RENDIMIENTO ESCOLAR	Insuficiente 4/10	Escaso 5/10	Suficiente 6/10	Bueno 7-8/10	Óptimo 9-10/10										

CUADRO 6.2. Modelo de aplicación de tareas (composición o escrito en torno a un tema)

a) Elemento que utiliza el alumno:

- Vocabulario específico (terminología)
- Variedad de frases
- Corrección de las frases
- Ortografía y puntuación
- Viveza de expresión
- Lenguaje matemático (complementario)
- Lenguaje plástico (complementario)

b) Contenido y estructuración:

- Señala la idea principal
- Señala las ideas secundarias
- Distingue la idea principal de las secundarias
- Presenta gran riqueza de ideas
- Explica la relación existente entre las ideas
- Distingue los hechos de las hipótesis u opiniones
- Utiliza datos reales (o documentos y pruebas)
- Hace una exposición ordenada
- Llega a conclusiones lógicas

c) Aspecto externo:

- Limpieza
- Gusto en la presentación

d) Errores:

- Número de errores
- Naturaleza de los errores (descripción)
- Posibles causas (descripción)

e) Originalidad:

- Hábil para enfocar muchas ideas a la vez
- Independencia de juicio
- Preferencia por lo complejo
- Autoafirmación
- Capacidad de síntesis
- Alto nivel de energía
- Pensamiento divergente

CUADRO 6.3. Modelo de análisis de una composición escrita

a) Contenido	1. Claridad de pensamiento 2. Continuidad de pensamiento 3. Sistematización de ideas
b) Elementos expresivos	4. Vocabulario 5. Variedad de frases 6. Viveza de expresión
c) Estructura	7. Corrección de las oraciones 8. División de frases o períodos dentro de la unidad general
d) Aspectos mecánicos	9. Ortografía 10. Puntuación

La utilización de escalas como las anteriores para valorar trabajos o tareas exige, por supuesto, que cada elemento o aspecto a valorar tenga asignado un peso o ponderación referente a la escala de puntuación (si la escala se puntúa); por ejemplo, de 0 a 50, habrá que asignar valores máximos o pesos a cada aspecto, de manera que esté ponderada su importancia y de manera, también, que la suma de los pesos sea el valor máximo de la escala, en este caso 50.

b)Las escalas de producción escolar persiguen el mismo objetivo, valorar o apreciar la calidad de trabajos presentados, pero a través de escalas graduadas u ordinales. Exigen un juicio por parte del evaluador, consistente en decidir el punto de la escala en que debe situarse el trabajo considerado. Son similares, en su estructura, a las escalas de estimación; lo que cambia es el objeto de estudio, en la escala de estimación se trataba de estimar conductas, aquí más bien aspectos o elementos de un trabajo escolar.

Al hablar de tests de rendimiento y en referencia a instrumentos externos, mencionaremos algunas baterías de pruebas pedagógicas en cuya calificación se utilizan escalas como las aquí mencionadas.

6.2. Instrumentos de encuesta

Los instrumentos de encuesta son, en muchas ocasiones, el procedimiento más adecuado para obtener muchas informaciones necesarias y relevantes al proceso evaluador. Pensemos en datos como los siguientes:

-¿Opinión del alumno sobre un asunto determinado?

-¿Intereses académicos?

-¿A quién admira un alumno?

-¿Cuál es el ambiente familiar de un alumno?

-Etc.

Muchas informaciones más, especialmente del ámbito afectivo, se pueden obtener rápidamente mediante los procedimientos de encuesta. Sin embargo, no debe olvidarse que la información recogida directamente de los alumnos puede ser muy subjetiva, pues todos los individuos tienden a dar respuestas "socialmente deseables", esto es, que les dejen en "buen lugar". De aquí que estas técnicas deban combinarse con datos de observación o de medida, sin duda mucho más fiables. Vamos a describir someramente los principales instrumentos de encuesta.

6.2.1. El cuestionario

Es una lista o repertorio de preguntas, formuladas por escrito, a las que el alumno puede responder. Las preguntas pueden ser cerradas o abiertas (elegir entre varias alternativas o dar una respuesta directa sin alternativas de elección). La administración o aplicación de los cuestionarios puede ser muy variada: en grupos pequeños o grandes y de forma presencial o bien por correo o por algún otro sistema que no exija realizarlo en presencia de los aplicadores. Algunos autores dividen a los cuestionarios en dos tipos: inventarios y escalas de actitudes. Nosotros, dadas sus características, incluiremos a estas últimas dentro de los tests y escalas de personalidad, intereses y actitudes.

Los cuestionarios, para que sean instrumentos útiles de recogida de información, deben intentar definir con precisión la información que es relevante, y formular cuestiones o preguntas dirigidas a recoger sólo esa información. En demasiadas ocasiones se observa que los cuestionarios construidos no han precisado claramente fines y objetivos, de aquí que incluyan cuestiones irrelevantes que plantean un doble problema: pueden, por su longitud y tiempo de aplicación, provocar cansancio en los encuestados, que tenderán a dar respuestas poco fiables y, por otro lado, a la hora de analizar la información, ésta es confusa y excesivamente prolija, siendo necesario seleccionar a posteriori la información relevante. Más grave aún puede ser que por falta de planificación del cuestionario, informaciones relevantes y necesarias no se recojan, anulándose o limitándose así resultados y conclusiones.

En consecuencia, un cuestionario deberá construirse para recoger información relevante de cara a una investigación o a un proceso evaluativo, siguiendo una serie de pasos lógicos:

a)Determinar con precisión qué datos son relevantes.

b)Formular cuestiones o preguntas sólo para recoger la información relevante.

c)Seleccionar, para cada información a recoger, el tipo de pregunta más adecuada (abierta o cerrada).

d)Evaluar cada cuestión en particular y el cuestionario en general, al objeto de garantizar - antes de su aplicación - que recogerá, en la forma más adecuada, la información que precisamos.

e)Evaluar también - antes de su aplicación - si el análisis posterior de la información podrá realizarse optimizando los resultados y conclusiones de la misma. No debe olvidarse que el análisis, cuantitativo o cualitativo, de los datos condicionará la forma de recoger, registrar y valorar la información.

6.2.2. La entrevista

La entrevista es, básicamente, un diálogo entre dos personas o entre una persona con un grupo; diálogo de carácter formal, serio, intencionado y con objetivos más o menos definidos.

La entrevista es, pues, un modo directo y personal de recoger información. Aquí radican sus ventajas y sus límites. Es una técnica que requiere una gran preparación, habilidad y experiencia por parte del entrevistador si desea conseguir, con fiabilidad y validez, datos relevantes de cara a cualquier proceso evaluador. Participa de las características de otros instrumentos tales como el cuestionario (habrá que planificar la entrevista formulando con precisión las cuestiones base para recoger información), los métodos de observación directa (se da una relación personal y directa entre entrevistador y entrevistado que permite anotar datos, circunstancias, comentarios e interpretaciones) y los propios tests (cuidado y precisión en la elaboración de ítems o reactivos de respuesta). He aquí su complejidad. No es sólo un instrumento o una técnica, sino un auténtico proceso de obtención de información. Por otra parte, sus modalidades pueden ser múltiples, haciéndola útil tanto para procesos de investigación como para procesos diagnósticos, de evaluación inicial y para procesos de orientación y consejo pedagógico o psicológico. Obviamente, en cada caso, presentará sus peculiaridades y matices, ya que los objetivos y fines de la entrevista condicionarán sus características y modalidad.

En este momento no vamos a desarrollar los múltiples aspectos que sobre la entrevista podrían considerarse. Sería demasiado complejo y extenso. Basta decir que hay numerosos tratados y libros sobre la misma. En todo caso, a través de entrevistas, informales o formales, el profesor puede obtener datos muy importantes para la evaluación. Sería adecuado que recibiese algún tipo de formación previa sobre la propia técnica, al objeto de que lograra familiarizarse con sus características, modalidades, posibilidades y limitaciones. Sólo así, la entrevista será utilizada como la poderosa herramienta informativa y formativa que en realidad es.

6.2.3. Técnicas sociométricas

Las técnicas sociométricas, cuyo origen debe buscarse en Jacob Lévy Moreno, son un conjunto de métodos e instrumentos cuyo objetivo es investigar la estructura social de un grupo. La sociometría, en su sentido más amplio, es la ciencia que trata de medir cualquier relación interpersonal; en su sentido estricto, se define por tratar de medir las relaciones preferenciales en una situación de elección. Moreno (1934) había definido previamente a la sociometría como "el estudio de la organización y evaluación de los grupos y la posición que en ellos ocupan los individuos, prescindiendo del problema de la estructura interna de cada individuo". Aun cuando esta última definición hoy no sea aceptada por la comunidad de sociómetras - ni siquiera Moreno (1957) la mantuvo en la segunda edición de su obra, Who shall survive?-, sin embargo, puede damos una idea muy cercana del conjunto de técnicas que incluye la sociometría al intentar estudiar y utilizar medidas sociales. Como afirma Pérez Juste (1983), gracias a la sociometría podemos conocer "si los sujetos son sinceros en sus respuestas, la auténtica estructura de los grupos, estructura que no siempre coincide ni en sus fines ni en su forma con la aparente o supuesta; este conocimiento es de gran utilidad cuando se trata de lograr eficacia en la consecución de metas de aprendizaje, especialmente cuando se utilizan los denominados equipos de trabajo, y en la solución de los problemas de adaptación de los escolares, pero también puede ser de gran valor para poner a prueba hipótesis científicas en la investigación educativa" (pág. 109).

El test sociométrico es, sin duda, la técnica sociométrica más conocida y utilizada y la de más interés para el educador o profesor, dado que siendo sencillo en su contenido y aplicación - no tanto en su interpretación - puede utilizarla para la recogida de información sobre su grupo de alumnos y sobre la peculiar situación de cada alumno en el grupo, lo que sin duda dará pistas importantes de cara al proceso evaluador en particular y al proceso educativo en general.

El test sociométrico"es el método empleado para conocer la estructura básica interrelacional de un grupo a través de las respuestas de sus componentes sobre sus propias atracciones y rechazos" (Arruga, 1987, pág. 21). Es, pues, un instrumento que mide la organización de los grupos sociales. Su contenido es muy simple: pedir a los sujetos de un grupo que expresen qué compañeros prefieren y qué compañeros les disgustan. Sobre la base de las atracciones y rechazos el sociómetra estudiará las estructuras sociales en su conjunto y la situación individual en que cada uno se encuentra.

Como vemos, el test sociométrico revela características tales como: subgrupos dentro de un grupo, "estrellas" o líderes, individuos rechazados o aislados, etc.

Existen distintos modelos de cuestionarios sociométricos. Tendrá que elegirse entre ellos, dependiendo del tipo de grupo (tamaño, edad, características) de que se trate y del

tipo de información que se desee recoger: criterios de elección o rechazo, ausencia o no y forma de las preguntas sobre rechazos, número de elecciones o rechazos, existencia o no de orden de preferencia y su ponderación cuantitativa, etc. De aquí que sea necesario introducirse en esta técnica para que sea utilizada adecuadamente.

Para la interpretación de los resultados se han hecho intentos como son la ponderación de las elecciones y rechazos recibidos, la elaboración de escalas centiles y la confección de determinados índices entre los que destaca el de estatus, el cual trata de apreciar el grado en que los sujetos conocen su situación en grupo, consistente en la relación o cociente entre las elecciones (o rechazos) que presume tener cada uno y las que realmente tiene (Pérez Juste, 1983).

En general, parece que la sociometría puede tener interesantes aplicaciones a la educación. De cara a la evaluación educativa, puede ser muy útil conocer, a través de los propios alumnos, las interacciones reales que se dan en una clase y la estructura social de la misma.

7

Instrumentos de medición: tests y pruebas

En este capítulo dedicaremos nuestra atención a los instrumentos de recogida de información que hemos clasificado como instrumentos de medición en sentido estricto. Comúnmente a estos instrumentos se les ha venido denominando tests. No vamos a entrar en este momento a discutir si una prueba objetiva de rendimiento o una escala de actitudes es verdaderamente o no un test. Nos interesa más una acepción amplia de dicho término que permita incluir instrumentos como los anteriores, esto es, instrumentos técnicamente construidos que permiten a un sujeto, en una situación definida (ante determinados reactivos o ítems), evidenciar la posesión de determinados conocimientos, habilidades, destrezas, nivel de logros, actitudes, características de personalidad etc. Son instrumentos que permiten apreciar una variable, tal como es definida por la misma prueba o instrumento.

Los tests, en el sentido amplio en que los venimos utilizando, a su vez pueden ser también divididos en internos y externos, con referencia explícita al lugar en que han sido construidos, dentro o fuera del centro escolar, respectivamente. En general, los tests externos, que se construyen fuera del centro escolar, suelen denominarse tests tipificados, y suelen ser elaborados por empresas u organizaciones privadas, estando destinados a medir ciertas características generales de los alumnos. Un test externo de medida debe atender a fines y criterios generales, porque su aplicación no va a estar restringida a una situación específica. Se les denomina tipificados o estandarizados, porque han sido normalizados o tipificados (de referencia normativa) los procedimientos de aplicación, de modo que todos los sujetos de un grupo normativo o población respondan a los mismos bajo idénticas condiciones de aplicación. En este sentido, el tiempo concedido para un test es el mismo para todos los examinandos y las instrucciones (leídas por el alumno o por el profesor en voz alta) son también idénticas en todas las ocasiones en que se aplica el test. El objetivo de normalizar los procedimientos es asegurar la comparabilidad de las puntuaciones que obtienen los alumnos en el test, porque han sido obtenidas bajo las mismas condiciones (Beggs y Lewis, 1979, pág. 88). Es obvio que, como explicamos al hablar de referencia a normas y a criterios, los tests tipificados pueden utilizarse para comparar la ejecución de un sujeto con la ejecución de un grupo cono cido (normativo), pero en ningún caso, las puntuaciones en dichos tests normativos permiten apreciar si dicha ejecución es satisfactoria o deseable.

Desde el punto de vista de la evaluación educativa, los tests externos (tipificados) que más nos interesan son los de aptitudes, actitudes y rendimiento, especialmente estos últimos. Los tests de personalidad, intereses, adaptación, etc., también tienen su importancia de cara a la evaluación educativa - especialmente en la evaluación inicial pueden ofrecer información relevante sobre características y posibilidades de los alumnos - pero dado su menor nivel de uso por parte del profesor les dedicaremos solamente una breve referencia.

Ahora bien, aunque los tests tipificados son ampliamente utilizados en nuestro país y en el mundo entero, no son sin embargo los más utilizados en los centros escolares. Ello es lógico, la mayoría de los tests usados en nuestros centros son elaborados por el profesor o equipo de profesores de una materia porque los instrumentos tipificados no evalúan conceptos cognitivos específicos implicados en las tareas diarias de los procesos de enseñanza-aprendizaje, sino más bien variables o aptitudes de tipo general. De aquí que la mayoría de los instrumentos utilizados en el proceso educativo sean medidas internas, que deben ser construidas y evaluadas con sumo cuidado. Por ello, dedicaremos nuestra atención preferente en esta obra a estos instrumentos internos de medida del rendimiento educativo.

Resumiendo, los tests se caracterizan por (Tenbrink, 1981):

a)Una situación estándar o común a la que todos los sujetos responden.

b)Unas instrucciones y tiempo comunes para todos los sujetos.

c)Un conjunto de normas o reglas comunes para corregir las respuestas de los sujetos.

d)Una descripción o resultado numérico de la realización de cada alumno, hecha después de corregir las respuestas.

Aceptando la división de los tests en externos (tipificados) o internos (construidos por el profesor o grupo de profesores de un centro), podemos caracterizar a los tests internos por estar dirigidos específicamente a medir el logro de los objetivos educativos definidos y a los tests externos tipificados por:

a)Medir rasgos o variables de carácter general.

b)Estar construidos por empresas privadas o especialistas.

c)Estar formados por ítems o tareas previamente experimentadas en grupos normativos. Garantías de fiabilidad y validez.

d)Aplicarse y corregirse estandarizadamente, esto es, de acuerdo a rigurosas instrucciones.

e)Su aplicabilidad restringida a grupos similares o de las mismas características al grupo normativo.

f)Estar generalmente comercializados.

Finalmente, es preciso recordar una recomendación general a la hora de seleccionar o construir instrumentos de recogida de información o medida: habrá de realizarse (tal selección o construcción) en función de los objetivos educativos de que se trate y de la infor mación que se considere relevante obtener en orden a la evaluación del proceso de enseñanza-aprendizaje del alumno.

7.1. Instrumentos externos de medida (tipificados)

Dentro de este grupo de instrumentos desarrollaremos brevemente a continuación tres tipos básicos:

a)De aptitudes.

b)De personalidad, intereses, actitudes, adaptación, etc.

c)De rendimiento o pedagógicos.

7.1.1. Tests tipificados de medición de aptitudes

Dichos tests pretenden medir habilidades innatas o aprendidas que se supone predicen facilidad y rapidez para el aprendizaje en una área determinada. Intentan, pues, medir potencial o capacidad de un sujeto de cara al éxito en tareas específicas. Ciertamente no son éstas las únicas definiciones posibles de "aptitud", pero sin entrar en la discusión sobre su definición, incluiremos bajo esta denominación a los tests de inteligencia general, de aptitudes mentales diferenciadas y de aptitudes especiales.

Algunos tests de aptitudes están diseñados para medir rasgos generales, mientras que otros están diseñados para medir rasgos específicos o muy específicos. La diferencia entre aptitud y rendimiento a veces no es clara, especialmente por el contenido de los tests. Sin embargo, una característica o factor que puede distinguirlos es el hecho de que los tests de aptitud se utilicen en la práctica como predictores, están diseñados para dar indicación de éxito futuro, mientras que los tests de rendimiento miden más bien éxito, aprovechamiento, logro. A pesar de esta diferencia entre ambos tipos de tests, bien es verdad que los propios resultados de un test de rendimiento pueden utilizarse como predicción de éxito futuro en tareas similares.

En principio, y además de la aptitud general o inteligencia general, existen tantas aptitudes como capacidades o habilidades puede teóricamente poseer un sujeto. De hecho, existen tests que miden gran número de estas aptitudes.

No nos vamos a detener en la explicación del cómo construir tests de aptitudes, baste decir que una vez definida conceptual y operativamente una aptitud, es preciso elaborar ítems que la estimen. Posteriormente, el test, en edición experimental, se pone a prueba en varios grupos de diferentes sujetos; se estudian los resultados psicométricamente (estudio de las características técnicas del test) a través de criterios estadísticos como fiabilidad, validez y análisis de los elementos. Depurada la prueba se procede a tipificar o establecer baremos para distintos grupos normativos en los cuales se supone que dicha prueba estimará fiable y válidamente el rasgo definido.

En un apartado posterior ofreceremos criterios para interpretar y valorar los resultados de los tests de aptitudes y en general las posibilidades y limitaciones de los tests tipificados.

De un modo orientativo, el siguiente cuadro muestra algunos de los tests de aptitudes más importantes empleados en nuestro país (cuadro 7.1):

CUADRO 7.1. Algunos instrumentos tipificados de medición de las aptitudes utilizados en nuestro país

1. Inteligencia general (Factor "G")

1.1. Base verbal

- Ballard
- CC-78. Factor IG de J. Crespo
- CM-76 de IG
- Otis

1.2. Mixtos

- Aptitudes generales inferior
- Escala de Wechsler WPPSI
- Escalas McCarthy
- WISC, escala de inteligencia
- WAIS, escala de inteligencia
- Escala de inteligencia de Zazzo

1.3. Base no verbal

Individuales

- Terman-Merrill
- Cubos de Binet
- Escala de madurez mental (Columbia)
- Escala de desarrollo del pensamiento lógico

Colectivos

- Beta
- D-40 y D-70 (Dominós)
- Figura humana (Goodenough)
- Factor "g" de Cattell
- Inteligencia general de Bonnardell
- Matrices progresivas de Raven (superior, inferior, en color)
- TIG-1 y TIG-2 (Dominós)

2. Aptitudes diferenciadas

2.1. Baterías de pruebas

- AMDI/AMPE/AMD-76
- DAT
- TEA-1, TEA-2 y TEA-3
- TCI

2.2. Razonamiento

- Inteligencia lógica
- Pruebas de razonamiento de G. Hoz
- Semejanzas y diferencias

2.3. Imaginación. Creatividad

- Asociación de palabras
- Creatividad de M. Beltrán
- AG 75 de Gosalvez

2.4. Aptitudes perceptivas y atención

- Percepción de Bonardell
- TP de Toulousse-Pieron
- Figuras idénticas de Thustone

2.5. Memoria

- AG-74 y AG-75 de Gosalvez
- Memoria de TEA

2.6. Destrezas manual y motora

- Aptitudes mecánicas de McQuarrie
- Aptitudes mecánicas de Inapp
- Varios tests de Bonnardell

2.7. Aptitudes espaciales

- Desarrollo de superficies (Yela)
- B-31, B-32, B-22 y B-43 de Bonnardell
- Test técnico práctico de Goguelin

2.8. Aptitudes verbales

- COE de TEA
- Compresión verbal (BV 17)
- Instrucciones complejas de Yela

2.9. Aptitudes verbales

- Aritmética (Pb 17), Bonnardell
- Cálculo (Add 15 y Mislt 16), Bonnardell
- Test numérico de TEA
- Rapidez de cálculo de INAPP

2.10. Aptitudes artísticas

- Apreciación de dibujos (Graves)
- Aptitudes musicales (Sehasshore y otros)

7.1.2. Tests de personalidad, intereses, adaptación, actitudes, estilos cognitivos

En general, los tests, inventarios, cuestionarios o escalas destinadas a medir estas variables, pretenden apreciar aspectos no intelectivos de la constitución mental o psicológica del individuo. En estos instrumentos, por lo general, no se trata de valorar las

respuestas predeterminadas para sus ítems. De aquí que en la mayoría de las ocasiones el término test no se utilice en su denominación estricta.

A) Tests de personalidad

La medición de la personalidad se ha contemplado fundamentalmente desde dos grandes perspectivas: aquella que la considera como única, global, total (perspectiva cualitativa que utiliza básicamente tests proyectivos) y aquella otra que trata de identificar aspectos, rasgos, variables, factores o dimensiones dentro de tal globalidad (perspectiva cuantitativa que utiliza básicamente tests psicométricos o normativos).

El primer enfoque, utiliza para apreciar la personalidad pruebas proyectivas integradas, por lo común, por elementos expresivos, cuya interpretación se realiza a través de informes cualitativos difícilmente tipificables y comparables; su uso fundamental se sitúa en el campo clínico o en estudios de casos. La mayoría de los expertos en medida opinan que son excesivamente subjetivos en su interpretación (no hay criterios unívocos) y que exigen una gran preparación y experiencia para aplicarlos e interpretarlos con ciertas garantías de fiabilidad.

El segundo enfoque, que básicamente está integrado por los inventarios, cuestionarios y escalas de personalidad, trata de apreciar rasgos o dimensiones de la misma, que sintetizan en una puntuación cuantitativa para cada rasgo o dimensión. Se supone que dichas variables son continuas, esto es, definen un contínuum de valores que oscila entre dos extremos o polos. La puntuación obtenida por un sujeto en una escala o test se sitúa en dicho continuo apreciando, en alguna manera, una tendencia del sujeto hacia uno u otro polo o extremo de la variable objeto de medida. La construcción e interpretación de estos instrumentos es similar a la esbozada al hablar de los tests de aptitudes. Son pruebas elaboradas con referencia normativa, tipificadas y que deben ser valoradas en términos de fiabilidad y validez psicométricas. Su problema básico es el de validez, dadas las dificultades de encontrar criterios con los que evaluarla (analizaremos este problema al hablar del estudio de las características técnicas de un instrumento de medida).

De modo orientativo, el cuadro 7.2 recoge los principales instrumentos utilizados en nuestro país para medir y apreciar variables de personalidad.

CUADRO 7.2. Principales tests psicométricos y proyectivos de personalidad

1. **Cuestionarios, inventarios o escalas.**

   - CAPPA. Cuestionario AP de Personalidad de Amurrio
   - Cuestionario de frustración DELTA-S (COSPA)
   - CEP de Pinillos
   - 16PF. Cuestionario factorial de personalidad de Cattell
   - PNP. Cuestionario de personalidad de Pichot
   - GF-5 y GF-10 de temperamento de INAPP
   - MMPI. Inventario multifásico de personalidad de Minnesota
   - EPI. Cuestionario de personalidad de Eysenk
   - Pruebas de personalidad de EOS
   - IP. Temperamento y carácter ("Saint Georges", Montreal)

2. **Pruebas proyectivas y clínicas.**

   - TAT. Test de apercepción temática de Murray
   - Test del Árbol de Koch
   - Test de la familia de Corman
   - Test de la figura humana de Machover
   - Test "Z" de Zullinguer
   - Test de las manchas de tinta de Holzman
   - Test de las manchas de tinta de Rorscharch

B) Inventarios de intereses

Un aspecto del individuo que también es importante estudiar, tanto para comprenderlo como persona, como para ayudarlo a resolver problemas prácticos o bien problemas relacionados con la orientación académica y profesional/vocacional, es el campo de los intereses y de las aversiones, de las preferencias y las cosas del medio que lo rodea. Por supuesto, y como afirman Thorndike y Hagen (1975), "en lo relativo a los intereses vocacionales, el procedimiento más sencillo consiste en preguntarle al individuo cuánto le gusta ser ingeniero, por ejemplo. Sin embargo, esto no sirve en la práctica. En primer lugar, la gente difiere por lo que toca a la forma en que exhibe su entusiasmo. "Me gusta mucho" en el caso de la persona A, quizá no signifique que tiene más entusiasmo que la persona B, cuando dice "me gusta". En segundo lugar, la gente difiere considerablemente por lo que se refiere a la naturaleza y al grado de su comprensión de lo que un trabajo particular significa en términos de actividades y condiciones de trabajo. "Ingeniero" para una persona puede significar primordialmente trabajo al aire libre; a otro le sonará a laboratorio o tablero de dibujo. De aquí, que estos significados variados, hacen que la respuesta a la pregunta: ¿cuánto le gustaría ser ingeniero?, sea un indicador más bien insatisfactorio del grado en que el individuo tiene intereses realmente adecuados para la profesión de ingeniero" (pág. 403).

Por estas razones, los constructores de tests y pruebas han ampliado la base de la

información a recoger y formulan las cuestiones en términos de preguntas acerca de "gustos" y "desagrados" del individuo o en vez de preguntas directas acerca de preferencias por oficios o empleos particulares.

Al igual que los tests de aptitudes, los inventarios de intereses son construidos normativamente, son tipificados para grupos o poblaciones particulares y valorados en términos de validez y fiabilidad psicométrica. También el problema fundamental de estas pruebas, como ocurre en general con el de todas las variables que tratan de medirse indirectamente, es el de su validez, en este caso, fundamentalmente predictiva del éxito en una elección profesional o de la satisfacción en el desempeño de una profesión. A continuación, mostramos un cuadro que recoge las principales pruebas de intereses utilizadas en nuestro país. Sin duda las de mayor tradición son el Inventario de Strong y el Registro de preferencias vocacionales de Kuder (cuadro 7.3):

CUADRO 7.3. Inventarios de intereses profesionales/vocacionales

- Aspiraciones TD De INAPP
- Intereses profesionales de Thurstone
- IOTA. Cuestionarios de intereses profesionales
- Orientación vocacional de Gosalvez
- Registro de preferencias vocacionales de Kuder
- Intereses vocacionales de Strong
- Temario vocacional TV de García Yagüe

C) Pruebas de adaptación

Otro aspecto importante del individuo que conviene estudiar es el de su adaptación, es decir, el de la relación del individuo consigo mismo y con su entorno, el de la adaptación entre su situación y sus deseos, entre su imagen y su realidad, etc. Obviamente, por definición, la adaptación nunca es total. Como afirma Pérez Juste (1976) "alcanzamos grados de mayor o menor satisfacción, niveles más o menos altos; por otra parte, alcanzados ciertos niveles satisfactorios, la situación no se hace perenne, sino que nuevos estímulos, situaciones diferentes, rompen el equilibrio logrado, desencadenan un nuevo dinamismo".

Así pues, la adaptación es un ií constructo o dimensión explicativa del comportamiento humano, muy relacionado con las dimensiones anteriormente estudiadas, que supone una relación entre su individuo y su entorno, su medio. El mismo autor, mencionado más arriba, clasifica estos criterios de la siguiente forma:

1. Tomada la adaptación desde el punto de vista estático:

   a) Consideración de los diversos seres con los que el sujeto puede entrar en relación.

b)Consideración de la relación que el sujeto establece consigo mismo o con el exterior:

-Zona subjetiva:

*Adaptación fisiológica.

Adaptaciónemocional.

-Zona objetiva:

*Adaptación familiar.

Adaptaciónescolar.

*Adaptación profesional.

*Adaptación social.

2.Considerada la adaptación desde el punto de vista evolutivo podríamos estudiar la adaptación a la infancia, a la adolescencia, a la juventud, a la madurez y a la vejez.

3.El tercer criterio sería la integración de los dos anteriores.

Las pruebas que tratan de estimar variables, áreas o dimensiones de adaptación han utilizado para su elaboración los criterios anteriores. Las principales pruebas de adaptación, de hecho, miden una o varias dimensiones y han sido construidas normativamente y tipificadas para distintos grupos o poblaciones. Por ejemplo, el Cuestionario para Adultos de Bell (adaptado a España por Cerdá) estima cinco dimensiones de adaptación:

a)Familiar.

b)Ala salud.

c)Social.

d)Emocional.

e)Profesional.

Además estima, a través de una puntuación total, una dimensión general de adaptación.

Las principales pruebas o instrumentos de adaptación utilizados en nuestro país son

los siguientes (cuadro 7.4):

CUADRO 7.4. Pruebas de adaptación

- Cuestionario de adaptación para adolescentes (Bell)
- Cuestionario de adaptación para adultos (Bell)
- RDI rasgos de identificación (inventario de adaptación del Instituto "Saint Georges" de Montreal)
- Cuestionario de adaptación social de Pérez Juste
- Cuestionario de adaptación escolar de Jiménez F.

D) Escalas de actitudes y de autoconcepto

Las escalas de actitudes pertenecen a los métodos que genéricamente se denominan de autoinforme, los individuos aportan información sobre sí mismos. Combinan las características de las escalas de estimación con las de los inventarios y en ellas se pide al sujeto que exprese su actitud con relación a una afirmación determinada, señalándola en una escala graduada, en un continuo definido por varios grados, generalmente más de dos y menos de ocho. Existen otros métodos o enfoques en la medición de las actitudes, como el recoger opiniones y valoraciones sobre conductas ajenas (heteroevaluación), los métodos basados en la observación directa de conductas, los métodos indirectos y no obstrusivos y otros métodos que aparecen sistematizados en numerosos textos y publicaciones. Al igual que los intereses y valores, las actitudes (como categorías de objetivos) se integran en el ámbito afectivo y por ellas se entiende: "predisposiciones aprendidas, no innatas, y estables, aunque pueden cambiar, a reaccionar de una manera valorativa, favorable o desfavorable ante un objeto (individuos, grupos, ideas, situaciones, etc.)" (Morales, 1987).

La medición de actitudes a través de escalas se basa en la medición de sus manifestaciones, que son reacciones valorativas ante opciones referidas a creencias (no conocimientos), sentimientos, o conductas.

Dada la variabilidad de posibles objetos hacia los que pueden manifestarse actitudes (sí mismo, individuos, grupos, ideas, creencias, situaciones, etc.) existe una gran variabilidad de escalas posibles y existentes para medirlas. En todo caso, dichas escalas podrían servir como modelo de construcción de otras escalas, pero puede no ser posible o no tener sentido aplicarlas, sin modificaciones, en otras situaciones o contextos distintos (sujetos o grupos o clase de actitud) a aquél en que una escala fue construida.

Las escalas de actitudes más importantes son las de tipo Likert, consistentes en una afirmación que el individuo debe valorar en términos de mayor o menor grado de acuerdo o mayor o menor grado de valoración positiva o negativa. Algunos ejemplos de ítems, tomados de Morales (1987) pueden ilustrar la técnica concreta tipo Likert de

medición de actitudes:

a)Actitudes religiosas

(–) Con tantas religiones diferentes uno no acaba de saber en qué creer.
(–) Lo importante en la vida es pasarla sin demasiadas penas.

5	4	3	2	1
Totalmente de acuerdo	De acuerdo	Indeciso	No estoy de acuerdo	No estoy de acuerdo en absoluto

b)Actitudes hacia la democracia

(-)Con democracia o sin ella siempre mandan los mismos: los más listos y los más poderosos.

(-)Los partidos políticos en general sólo se preocupan de verdad de proteger los intereses de unos pocos.

(-)Las decisiones del Gobierno son como el tiempo: no se pueden predecir.

5	4	3	2	1
Totalmente de acuerdo	De acuerdo	Indeciso	No estoy de acuerdo	No estoy de acuerdo en absoluto

c)Actitudes hacia la conservación de la naturaleza

(-)Los lobos están bien en lugares acotados, pero no donde pueden causar daños al ganado.

(-)No debe detenerse el progreso de una comarca con la excusa de proteger a los pájaros.

(-)No parece muy ético preocuparse tanto de los animales salvajes cuando hay tantos seres humanos que lo pasan muy mal.

(-)Está bien conservar animales salvajes, solamente si producen algún tipo de beneficio económico.

d)Escala de autoconcepto (actitud hacia sí mismo)

(-)Me gustaría ser distinto de lo que soy.

(-)Cedo con facilidad ante los demás.

(-)A veces me gustaría marcharme de casa.

(+)Estoy seguro de que las cosas me irán bien en la vida.

(+)Soy una persona divertida.

(-)A veces me parece que no sirvo para nada.

(+)Yo creo que caigo bien a mis compañeros.

e)Escala de actitudes hacia el estudio

(+) Para mí es muy importante sacar notas altas.

Sí, mucho	Bastante	Más bien no	Ciertamente no

(-) Estudiar es duro y aburrido.

Sí, mucho	Bastante	Más bien no	Ciertamente no

(+)Me esfuerzo en comprender las cosas difíciles.

Sí, siempre	Sí, pero no mucho	Más bien poco	Muy poco o nada

(+)Me gustaría saber más sobre muchas de las cosas que estudiamos en clase.

Sí, mucho	Sí, pero no mucho	Más bien no	Por supuesto que no

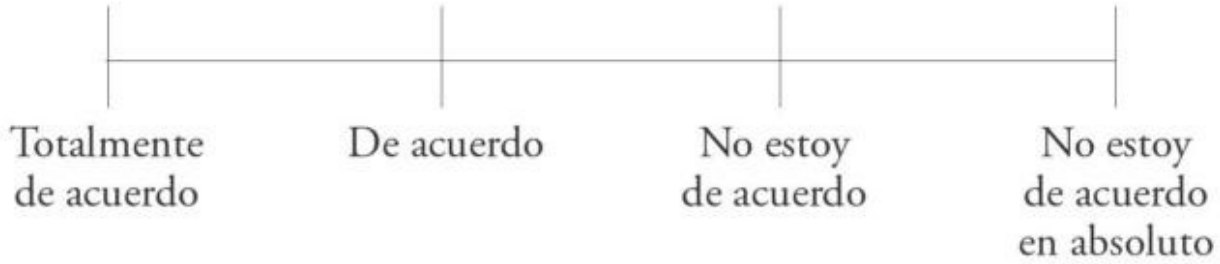
f)Escala de motivación de logro (búsqueda de éxito, competitividad, aceptación de riesgos, constancia y organización del trabajo, proponerse objetivos a largo plazo)

(-)Prefiero hacer algo en lo que me encuentro seguro y relajado que meterme en algo más dificil y que es para mí como un desafio.

(+)Preferiría llevar mi propio negocio y arriesgarme a un 50 por 100 de probabilidades de terminar en bancarrota, que trabajar para otro.

(-)Pienso más en mis éxitos pasados que en mis metas futuras.

(+)Prefiero trabajar a comisión, con más riesgo pero con posibilidades de ganar mucho dinero.



g)Actitudes éticas

	1ª Escogería (entre A y B)	2ª De mi elección estoy
1. Halagar e intentar hacerme amigo de quien me puede ayudar	(A)	( ) Muy seguro
2. Gastar tiempo y energías en ayudar a personas que no pueden favorecerme en nada	(B)	( ) Poco seguro
3. Callar cosas que son verdaderas pero me podrían perjudicar	(A)	( ) Muy seguro
4. Callar cosas que son verdaderas pero que podrían dar ventaja a mi competidor	(B)	( ) Poco seguro

Como vemos, los ítems pueden tener más o menos alternativas, y formularse en términos positivos o negativos. Normalmente se puntúa, en estas escalas sumativas, asignando valores a cada respuesta del sujeto (entre 1 y 5 en una escala de cinco alternativas; o entre 0 y 1, si sólo hubiere dos) y sumando las puntuaciones obtenidas en cada ítem de la prueba. Se obtiene así una puntuación global que expresa grado de actitud positiva hacia el "objeto" de que se trate, una puntuación elevada indica alto grado de actitud positiva y una puntuación baja indica actitud negativa o baja actitud positiva hacia el "objeto" o variable implicada en la definición de la escala. Si el objeto de medida es la actitud hacia uno mismo, la escala mide lo que suele denominarse autoconcepto o autoestima. No obstante, el estudio de las actitudes, opiniones y estimación de uno

mismo constituye un área específica, a la que se suele denominar autoestima y autoconcepto, no siendo ambas variables la misma cosa. Por lo general, la mayoría de los autores piensan que el constructo o dimensión de autoconcepto es más general que el de autoestima, pero no vamos a entrar en este momento en la discusión de este tema.

E) Tests de estilos cognitivos

A principios de los años cincuenta se empezaron a estudiar los que posteriormente fueron denominados estilos cognitivos. Dichos estilos, nacen al intentar cubrir un hueco que existía entre los factores puramente cognitivos y los factores de personalidad, en un intento de ahondar en el problema de las diferencias individuales. Como dimensión psicológica pronto adquirieron un gran auge, ya que se vio como una alternativa al CI (cociente intelectual) al no parecer que éste fuera un índice de predicción de ciertas actividades cognitivas (creatividad, empatía, etc.).

Los investigadores de los estilos cognitivos indican que hay dos tipos básicos de pensamiento: el analítico y el integrativo (o globalista). El CI sólo cubría el analítico. El interés por la temática cognitiva en los últimos treinta y cinco años se ha centrado en dos dimensiones:

a)La búsqueda de los principios y leyes generales de funcionamiento cognitivo. Es la más conocida y relevante para la psicología básica.

b)El análisis de ese funcionamiento en tareas concretas. Esto es, búsqueda del modelo de funcionamiento cognitivo, así como en sus posibilidades predictivas y su posible modificación. Estos estudios han permitido encontrar importantes diferencias individuales a la hora de enfrentarse con un mismo problema. Estas diferencias tenían más que ver con la forma, la manera, el cómo, que con el resultado final o la eficacia.

Así pues, los estilos cognitivos se hallan en la bisagra entre las dimensiones estrictamente cognitivas y los distintos aspectos de la personalidad (aunque esto es lo que les hace más atractivos, también es lo que más complica su conceptualización a la hora de intentar operativizarlos y controlarlos).

A pesar de la multiplicidad de definiciones que sobre estilos cognitivos se han dado, es posible distinguir dos enfoques o grandes grupos de definiciones:

a)Los que resaltan el carácter fronterizo de los estilos cognitivos, es decir, los que resaltan el hecho de que no sean dimensiones puramente cognitivas, sino que están relacionados con aspectos de la personalidad (Witkin, 1976 y 1977; Kogan, 1976, etc.).

b)Los que ponen su interés esencial sobre aspectos cognitivos, sobre las diferencias entre sujetos en lo relativo a las estrategias y procedimientos de que se sirven en la resolución de problemas (Kogan, Moss y Sigel, 1980; Sigel y Coop, 1980; Cohen, 1969; Goldstein y Blackman, 1978, etc.).

En resumen, los estilos cognitivos son constructos, dimensiones sintetizadoras que se han venido definiendo para resumir el método cognitivo básico en que una persona se apoya al enfrentarse a todas las formas de estimulación sensorial. Tratan, pues, de definir modos generales de cada individuo de procesar la información que percibe del ambiente y de enfrentarse a situaciones en las cuales debe desarrollar un procedimiento y estrategia de resolución.

Las notas fundamentales que les caracterizan y distinguen de otras variables son, en resumen, y siguiendo básicamente a Entwistle, E. (1985):

a)Hacen referencia a las diferencias entre sujetos en términos de cómo perciben, piensan, solucionan problemas, se relacionan con ellos (se centran más en la forma que en el contenido de la actividad congnitiva).

b)Son dimensiones generales y penetrantes. Esto hace que traspasen los límites de lo puramente cognitivo de la persona. De aquí que permitan hablar de la persona en su carácter total (holístico) y valorar los atributos denominados "no cognitivos".

c)Son estables y persistentes en el tiempo, lo que no implica que no se puedan modificar en ningún estadio evolutivo. No obstante, permiten predicciones con cierta seguridad, lo que hace que los estilos cognitivos sean dimensiones particularmente útiles para la orientación y asesoramiento a largo plazo.

d)Son bipolares, no teniendo - en principio - ningún polo de connotaciones negativas, sino sólo valores diferenciados. Cada polo representa una capacidad adaptativa distinta. Esto es, son variables de carácter neutro.

Existen definidas infinidad de distintas dicotomías pertenecientes a otros tantos estilos cognitivos. A la hora de presentar una clasificación de los mismos parece adecuado presentarlos en términos de:

a)Estilos de percepción.

b)Estilos de procesamiento de la información.

c)Estilos de pensamiento.

d)Estilos de aprendizaje.

Esta clasificación atiende a la naturaleza de la variable predominante en ellos, así como a la misma medida de dichos estilos. El cuadro 7.5 permite recoger a las dimensiones más importantes, así como las diferentes categorías de sujetos que de su consideración se derivan.

CUADRO 7.5. Principales dimensiones de estilo cognitivo (Entwistle, 1985)

Tipo de estilo	*A + B Integrado*	*Estilo A*	*Estilo B*	*Subdesarrollo*	*Autores*
Percepción	– Rápido-exacto	Dependencia de campo impulsivo	Independencia de campo Reflexivo	Lento-impreciso	Witkin Kagan
Procesamiento de la información – Categorías – Conceptos	– –	Amplio Relacional	Racional Analítico	– Descriptivo	Wallach Kogan
Pensamiento	Ambos altos	Divergente (Imaginativo)	Convergente (lógico)	Ambos flojos	Hudson
Aprendizaje	Versátil	Holístico	Serialista	Rutinario	Pask
Personalidad	Integrado	– Expresivo – Emocional – Complejidad cognitiva	– Autocontrol – Inhibido – Práctico	– – –	Hudson Entwistle Morrison

De todos los estilos cognitivos, los más estudiados en los últimos años y especialmente en nuestro país han sido las dimensiones dependencia-independencia de campo perceptivo de Witkin e impulsividad-reflexividad de Kagan. Actualmente sólo la dimensión definida por Witkin se halla comercializada (TEA), aunque también, en los departamentos universitarios de investigación educativa y psicológica, se disponga de otros instrumentos de medida de estilos cognitivos para su estudio, valoración y utilización en investigación.

7.1.3. Tests e instrumentos externos de medición del rendimiento y de otras áreas pedagógicas

Un test de rendimiento está construido para estimar o medir el grado o nivel en el que un sujeto domina o posee ciertos conceptos y capacidades que corresponden a un contenido específico y en un determinado momento de su desarrollo evolutivo. Por ejemplo, un test de lectura podría intentar medir o determinar el conocimiento que un alumno tiene de las palabras, velocidad, fluidez, o su capacidad para distinguir los signos de puntuación en una redacción. Un test de cálculo permitirá determinar el grado en que un alumno es capaz de realizar operaciones básicas como sumar, restar, multiplicar, dividir o bien resolver problemas enunciados verbalmente.

Como es lógico, los tests externos de medida del rendimiento están construidos para diferentes niveles de edad, dado que no puede esperarse que sujetos de diferentes cursos y edades posean las mismas capacidades. Así pues, como afirman Beggs y Lewis (1979), "la finalidad de los tests de rendimiento es definir el grado en que un alumno posee los con ceptos y capacidades que son razonablemente esperados en ese momento del desarrollo del sujeto o fase de estudio, o aquellos que son, generalmente, aceptados como importantes por los especialistas que elaboran el test de rendimiento" (pág. 90). La interpretación de las puntuaciones obtenidas en dichos tests tipificados se basa generalmente en el sistema de referencia a normas o baremos, es decir, los resultados de un alumno en el test se comparan con la ejecución alcanzada por aquellos alumnos que integran el grupo normativo al cual se supone que pertenece.

Es evidente, aunque no todo el mundo esté de acuerdo, que la información obtenida por un profesor utilizando dichas pruebas, es valiosa, al permitirle comparar los resultados de un alumno particular con los de un grupo normativo, dado que le permitirá conocer en qué medida dicho alumno posee y domina los conocimientos, habilidades o destrezas propias de un estadio evolutivo determinado o propios de un grupo de referencia comparable. Por otra parte, dicha información puede también serle útil para valorar la eficacia de su enseñanza, así como ofrecerle información sobre el progreso individual de un sujeto de una aplicación a la siguiente y sobre diferentes subtests comparados entre sí. Incluso, como afirman Beggs y Lewis (1979) "el profesor puede tener alguna indicación sobre los resultados que se pueden esperar del sujeto en el futuro" (pág. 90); esto es, puede servir de base para la predicción del éxito futuro en contenidos o áreas similares o relacionadas con aquella que fue objeto de medida a través del test tipificado.

¿Cómo se construye un test tipificado de rendimiento? Generalmente las casas comerciales que los publican ofrecen esta información, pero no existe un formato definido que determine qué presentar y cómo presentarlo. A modo de esquema, esbozaremos brevemente los pasos de su construcción.

a)Determinación del área de contenido

Identificar el área de contenido a cubrir y la naturaleza del grupo de sujetos para quienes se va a construir. Debe ponderarse la importancia relativa de cada una de las subáreas de contenido implicadas.

b)Definición de los objetivos a conseguir para cada unidad de contenido

Para este paso las taxonomías de objetivos, como las descritas anteriormente, son de gran ayuda (Bloom, Gagné, Kratworld, etc.): conocimiento, comprensión, aplicación, análisis...

Suele construirse una tabla de doble entrada, denominada comúnmente tabla de especificaciones (véase capítulo anterior dedicado a la medida en la evaluación educativa). En ella se ponen en relación objetivos y contenidos; definiendo así qué objetivos deben cubrir los distintos contenidos, o lo que es lo mismo, con qué contenidos se piensa conseguir los objetivos propuestos.

c)Construcción de los elementos o ítems del test

A continuación, los constructores de tests elaboran elementos potenciales del test. Habrá que seleccionar el formato; normalmente suelen ser elementos de elección múltiple, dado que su corrección se considera más fácil y objetiva. Lógicamen te la redacción de elementos debe tener validez de contenido, esto es, adaptarse adecuadamente al objetivo-contenido a medir. Para ello, no sólo se tendrá en cuenta el contenido, sino también los distintos procesos cognitivos implícitos.

d)Puesta a prueba de los elementos del test

Con los elementos redactados se organiza provisionalmente la estructura de la prueba en la que debe estar ponderada la importancia dada por el constructor a determinadas subáreas de contenido y a determinados procesos o niveles cognitivos.

La edición experimental de la prueba se aplica a grupos de sujetos de las características y naturaleza especificadas en el primer paso, esto es, con las experiencias previas definidas en los prerrequisitos del test. También se aplica dicha prueba a grupos con niveles superiores o inferiores a los prerrequisitos, al objeto de valorar si cada ítem es adecuado realmente para el nivel para el que fue elaborado.

A continuación, se realiza un análisis de las características técnicas de cada elemento (cálculo de índice de dificultad y de discriminación) al objeto de evaluar la idoneidad de los índices en los diferentes niveles.

e)Elaboración de la prueba definitiva

Con los datos obtenidos en el paso anterior se redacta la prueba "definitiva", definiendo el orden de los elementos, el tiempo de aplicación, las instrucciones de aplicación, etc.

Normalmente esta prueba "definitiva" se somete a ulteriores estudios psicométricos que tienen como objetivo evaluar la fiabilidad y validez de la prueba para los diferentes grupos normativos o subpoblaciones a los que va dirigida.

Posteriormente, se suele baremar el test en dichos grupos, con el objeto de facilitar la comparación de las puntuaciones de sujetos medidos con esta prueba y la interpretación de las puntuaciones. Dicha baremación o tipificación suele realizarse en

escalas tipificadas o puntuaciones cuantiles (pentas, eneatipos, puntuaciones típicas, puntuaciones derivadas, centiles, deciles, cuartiles, etc.), para facilitar la interpretación de los resultados obtenidos al aplicar la prueba.

f)Manual del test

Finalmente, el constructor del test debe preparar un manual del test y aquellos otros materiales auxiliares para los usuarios, de modo que estos dispongan de información suficiente para determinar si el test es adecuado a sus propósitos y los resultados pueden ser utilizados correctamente. Existe una normativa fijada al efecto. Dicho manual debe contener, al menos, información sobre: interpretación de resultados (indicando qué escalas se aconsejan para minimizar los riesgos de interpretación incorrecta); validez del test (predictiva, concurrente...); fiabilidad del test (estabilidad, homogeneidad, consistencia interna de la prueba); instrucciones de aplicación, corrección y puntuación; baremos apropiados para diferentes grupos o subpoblaciones.

Aun cuando no existen en nuestro país muchas pruebas estandarizadas para medir el rendimiento académico en distintas áreas de contenido y edades, parece adecuado incluir en este momento algunos de los más importantes que existen. Algunos de los que mencionamos en el cuadro 7.6 no son propiamente tests tipificados de rendimiento, sino más bien pruebas, más o menos estandarizadas, que fueron construidas con el objetivo de medir variables pedagógicas en general, o de rendimiento en algún área básica en particular. Se trata de pruebas de lectura, escritura, cálculo, ortografía, composición escrita, técnicas y hábitos de estudio, etc. Algunas de ellas, quizás no son adecuadas para ser utilizadas en la actualidad - no han sido revisadas recientemente-; no obstante, las incluimos a modo de referencia para orientar al lector sobre el tipo de pruebas que han sido construidas en nuestro país (cuadro 7.6).

CUADRO 7.6. Algunos tests y pruebas estandarizadas de rendimiento y otras áreas pedagógicas

1. Baterías de pruebas pedagógicas
   - Conceptos básicos de Bohem
   - Pruebas pedagógicas de García Yagüe
   - Escala de instrucción de García Hoz
2. Lectura
   - ABC de Lourenso Filho
   - Batería de lectura de Inizan
   - Prueba de comprensión lectora de Ángel Lázaro
   - Dislexia de EOS
   - Prueba de lectura de Pérez González
   - Lectura oral de García Hoz
   - Prueba de velocidad lectora de Ángel Lázaro
   - Vocabulario de Salvador y Plantada
3. Escritura
   - Escala de escritura de Fernández Huerta
4. Composición escrita
   - Escala de composición de Gali
   - Evaluación de la composición escrita de Pérez González
5. Ortografía
   - Diagnóstico ortográfico analítico de Pérez González
   - Escala de dificultades ortográficas de Pacheco
   - Tests de ortografía de uso y de ortografía de reglas
   - Ortografía de TEA
6. Matemáticas
   - Aritmética de García Yagüe
   - Aritmética mecánica de Ballard- García Hoz
   - Cálculo N-1 de INNAP
   - Razonamiento aritmético de Ballard-García Hoz-Fernández Huerta
   - Prueba de iniciación al cálculo, escritura de números, operaciones y problemas
7. Técnicas y hábitos de estudio
   - ALFA de Cospa/IHE de Fernández Huerta
   - C. de hábitos de estudio de Illueca
   - ACH de Caballero
8. Dibujo
   - Escala de dibujo de García Hoz

7.1.4. Valoración e interpretación de los tests tipificados

Cuando es preciso elegir un test tipificado de aptitudes, o un test tipificado de comprensión lectora, por ejemplo, la cuestión que surge primero es: ¿cómo decidirse por uno entre los disponibles? La respuesta a esta cuestión se da al valorar las características técnicas de los distintos instrumentos de medida y su idoneidad para la situación en que va a ser aplicado, en términos de: fiabilidad, validez y utilidad de las puntuaciones que ofrece (en la Teoría Clásica de Test).

Cuando un constructor de pruebas referidas a normas (tipificadas) elabora un test, suele decirse que debe evaluar al mismo en términos de: fiabilidad, validez y características técnicas de los ítems; a la par debe ofrecer baremos del test que permitan comparar puntuaciones de sujetos de las mismas características a las del grupo normativo utilizado en la tipificación o, al menos, ofrecer indicaciones sobre el tipo de escalas más adecuado para tipificar la prueba en distintos grupos normativos.

Cuando un profesor decide seleccionar un test, de entre varios disponibles, debe sobre todo valorar su fiabilidad, su validez y la utilidad de las puntuaciones que él ofrece de acuerdo a sus objetivos. No le interesa, en principio, valorar las características técnicas de los ítems del mismo.

A) Fiabilidad

La fiabilidad de un test informa al usuario del mismo sobre la consistencia de las puntuaciones obtenidas, sea cual sea la variable medida. La fiabilidad suele estimarse a través de un índice que informa sobre la precisión o consistencia de los resultados obtenidos - nada dice este índice sobre la materia o contenido que pretende estimar el test - cuando los mismos individuos son examinados:

a)En diferentes ocasiones.

b)Con conjuntos distintos de elementos equivalentes.

c)Bajo otras condiciones variables de examen.

Este concepto de fiabilidad es la base del cálculo del error de medida de una puntuación sencilla, por el que podemos predecir la amplitud probable de la fluctuación en la puntuación de un sólo individuo, como resultado de factores causales.

Toda medida de la fiabilidad de un test informa sobre qué parte de la varianza total de las puntuaciones es varianza debida al azar, a errores, o a la casualidad. En esta varianza de error influyen todas aquellas condiciones que no sean afines al propósito de la prueba.

La fiabilidad, pues, alude a la concordancia o al grado de consistencia entre dos tipos o conjuntos de puntuaciones derivados de forma independiente (dos series de valores correspondientes a un mismo grupo de sujetos); por ello, puede expresarse siempre en función de un coeficiente de correlación. Este coeficiente, por definición y en esencia, expresa un grado de correspondencia o relación entre dos series de valores. Oscila entre - 1 y + 1, indicando el valor + 1 correlación perfecta positiva, el valor o ausencia de relación y el valor - 1 correlación perfecta negativa. De hecho pueden estimarse índices de fiabilidad a partir de distintos procedimientos:

a)Fiabilidad como estabilidad

Consistencia de las puntuaciones del test, estimada a través de la correlación entre técnica estadística utilizada en esta modalidad, llamada test-retest, es el coeficiente de correlación hallado entre ambas medidas de los mismos sujetos. Cuanto más se acerque a 1,00, mayor estabilidad, mayor fiabilidad tendrá la prueba. Ello indicaría que los sujetos que puntuaron más alto en la primera aplicación también fueron los que puntuaron más alto en la segunda y lo mismo con los que puntuaron con valores medios o bajos. No significa que obtuvieron la misma puntuación en ambas medidas, sino que mantuvieron el orden de puntuación entre los sujetos de medida a medida.

b)Fiabilidad como equivalencia

Consistencia de las puntuaciones del test, estimada a través de la correlación entre dos formas paralelas o equivalentes - no iguales - del mismo rasgo o variable a medir. El tipo de fiabilidad estimada es una valoración en los sujetos de la consistencia entre dos formas equivalentes de un test aplicado en dos momentos distintos. Cuanto más se acerque a 1,00, mayor equivalencia de ambas formas y mayor fiabilidad del test. Si el potencial usuario sólo desea un instrumento de medida externa para medir una sola vez a los sujetos, no es muy útil elegir este procedimiento para valorar la fiabilidad.

c)Fiabilidad como homogeneidad o como consistencia interna

Cuando todos los ítems o elementos de un test miden el mismo rasgo, tiene sentido evaluar su fiabilidad en términos de homogeneidad o consistencia interna. Un procedimiento usual consiste en dividir la prueba en dos mitades de elementos o ítems y correlacionarlas. El coeficiente obtenido se corrige a través de alguna fórmula estadística, como la de Spearman-Brown, y tal valor es el índice de homogeneidad, de fiabilidad de la prueba. Otro procedimiento consiste en analizar las correlaciones entre ítems resumiéndolas en un índice de consistencia interna o fiabilidad (coeficiente Alpha de Cronbach o fórmulas de Kuder-Richardson).

Cuando un docente quiere seleccionar un apropiado test tipificado, debe preocuparse por su fiabilidad y por su validez. Sin duda, como veremos, la validez es lo más importante. De la fiabilidad suele decirse que es condición necesaria, aunque no suficiente de un buen instrumento de medida. Por muy válido que hipotéticamente fuera, si no es fiable, si no es preciso, si no es consistente, las estimaciones del rasgo a través de las puntuaciones del test carecerían de valor dado que el error en dichas estimaciones sería muy elevado.

B) Validez

Como ya hemos dicho, es la característica más importante de un test. Una medida externa o test tipificado será válido si mide aquel aspecto para el cual fue diseñado o construido. En general, la validez de un método es la exactitud con que pueden hacerse medidas significa tivas y adecuadas con él, en el sentido de que midan realmente los rasgos o variables que pretenden medir. Es evidente, pues, que cuando queremos estimar la validez de un instrumento de medida, necesitamos saber qué rasgo deseamos que mida. A este rasgo suele denominársele variable de criterio. Nos interesa saber en qué grado se corresponden las posiciones de los individuos en la distribución de las puntuaciones obtenidas en el test a validar y los datos que usamos, con un grado conocido de certeza, como índices para las puntuaciones de los individuos en la variable criterio.

El tema de la validez de un test es ciertamente complejo. Son muchas las dificultades

que suelen encontrarse cuando queremos estudiar cualquier rasgo o variable. Vamos a comentar brevemente algunos de estos problemas. Un primer problema es el que se deriva del carácter indirecto de la medición, ya que nunca medimos un rasgo, sino algunas de sus manifestaciones conductuales, las cuales, suponemos, nos permiten estimar el grado en que un individuo "posee" un rasgo o variable. En este sentido nunca podrá decirse que un test es válido o no, sino en qué grado es válido para medir lo que pretende medir. Una restricción más, cuando pretendemos asignar un grado de validez a un instrumento, es la población en que dicha validez es estimada, dado que un instrumento podrá ser válido (en un cierto grado) para una determinada población de sujetos y no ser válido (en el mismo grado) para otra población de sujetos distinta.

De lo anterior se deduce claramente que la validez no es un concepto único, sino que adopta muy variadas formas e interpretaciones dependiendo del tipo de rasgo a medir, de los sujetos a que se destina, del uso que se hará de las puntuaciones del test, etc. De hecho, existen distintos tipos de validez, según la terminología usada por los autores que la definen; las más comúnmente aceptadas son: validez aparente, validez de contenido, validez predictiva, validez concurrente, validez de constructo y validez de utilidad; si revisamos la bibliografía sobre el tema, veremos que otros adjetivos se unen al sustantivo "validez" (tales como: incremental, convergente, discriminante, factorial, etc.), pero la mayoría de estos adjetivos quieren significar algunos aspectos esenciales que cada autor señala dentro de la validez de constructo.

En todo caso, las principales cuestiones a que pretende responder el estudio de la validez podrían sintetizarse de la forma en que lo hacen Martínez Arias (1981) y Cronbach (1971):

-Adecuación con lo que mide la conducta de una persona.

-Capacidad de diagnosticar la conducta de una persona.

-Sensibilidad para detectar un objetivo.

-Seguridad en las decisiones.

-Utilidad.

En esta línea, algunas conclusiones que pueden extraerse de todo lo dicho son (García Ramos, 1986):

La validez no es un concepto único; adopta diversas y variadas formas e interpretaciones según sea la variable a validar, los objetivos del test, la población a que se aplique dicho instrumento, etc.

-Lo que se valida no es estrictamente el instrumento, sino la interpretación de los datos

obtenidos por un procedimiento específico (Cronbach, 1971).

-La validez no debe entenderse como el cálculo de un simple índice que indique la relación entre test y criterio, sino más bien como un proceso continuo (Nunnally, 1978) que requiere la acumulación de datos procedentes de distintas investigaciones empíricas y de procesos lógicos.

-La validez no puede determinarse en términos absolutos respecto de ninguna variable de nuestro campo de estudio, sino que es una cuestión de grado.

-La elección de uno u otro tipo de validez o de varios tipos a la vez dependerá de las funciones típicas u objetivo que cumplirá el instrumento de medición particular que sea estudiado.

A modo de resumen, vamos a comentar, brevemente, los cuatro tipos de validez más importantes.

1. Validez de contenido

Este tipo de validez implica la adecuación con que los elementos de un test miden un contenido claramente especificado. Se determina revisando sistemática y lógicamente los elementos que forman parte del test, intentando comprobar si efectivamente dichos elementos permiten describir el área de contenidos-objetivos que definen la conducta a medir. Obviamente, exige la definición previa de tales objetivos en referencia a un contenido delimitado. Este tipo de validez lógica es especialmente útil para valorar un test o prueba de rendimiento, dado que en este tipo de test es posible delimitar con cierta precisión el dominio de conductas que el test debe intentar apreciar. Normalmente las casas editoras de tests ofrecen información a los potenciales usuarios sobre el contenido que cubren los tests, lo cual facilita la elección, de entre varios, conforme a la situación específica en que serán utilizados. En todo caso, y como afirma Pérez Juste (1983), "la esencia de la validez de contenido se encuentra en la suficiencia y representatividad de la muestra de conductas incluidas en el instrumento, respecto de la población de que forma parte" (pág. 227). Es claro que, siempre que sea posible, un instrumento de medida debe tener un número suficiente de elementos que permita apreciar el universo de conductas - esto es, que represente bien - que definen el rasgo o variable a medir.

2. Validez predictiva

Este concepto hace referencia a la relación existente entre las puntuaciones obtenidas en el test construido y alguna variable evaluada o evaluable posteriormente. Esto es, alude a las posibilidades que el test ofrece para predecir, estimar, avanzar juicios de cara al futuro. Este tipo de validez suele evaluarse a través de la correlación entre el test (predictor) y un criterio (variable a predecir), medidos ambos en una muestra de sujetos

correspondiente a una población bien delimitada. Obviamente, este tipo de validez, para su consideración, exige encontrar variables-criterio relevantes que permitan estimarla (criterios intermedios o finales). Se aconseja estimar la validez predictiva junto a la validez de contenido.

3. Validez concurrente

Concepto similar al anterior. También se estima correlacionando test y criterio establecido. La diferencia entre estos dos tipos de validez está en que en el caso de la validez concurrente el criterio ha de ser una variable contemporánea a la medida al test a validar. En ocasiones, por ejemplo, se utilizan tests ya validados (con garantías de validez, fiabilidad, etc.) como criterios para estimar la validez concurrente de otros tests construidos para medir la misma variable o rasgo.

4. Validez de constructo

El más importante tipo de validez, desde el punto de vista científico, según la mayoría de los autores, ya que profundiza al máximo en el tema de la validez, al intentar no sólo validar el test, sino también el rasgo o constructo que bajo él subyace. Este tipo de validez, definido en 1954 por la American Psychological Association, se propuso para "situaciones en las que el constructor del test no tenga una medida definitiva de criterio ante la cualidad o rasgo que le interesa apreciar y deba usar medidas indirectas para validar la teoría" (pág. 214). La validez de constructo, pues, permitirá que nos acerquemos a la medida de variables derivadas de una teoría concreta, sea ésta una teoría compleja o bien una serie de proposiciones simples. Ciertamente, los constructos o variables de construcción serán las que determinarán qué conductas han de ser seleccionadas para su observación y la situación en que será aplicada. Un instrumento de medida estará, pues, ligado al sistema de constructos dentro del que fue construido y es este tipo de validez el que nos indicará el grado en que el instrumento de evaluación es una medida adecuada del constructo y en qué medida las hipótesis derivadas del mismo pueden confirmarse mediante la utilización del instrumento (García Ramos, 1986).

C) Utilidad

Siendo la validez la característica más importante de un test y, como afirman Beggs y Lewis (1979), "es concebible un test perfectamente válido cuyas puntuaciones no sean útiles". De aquí se deriva que los tests, además de fiables y válidos, deben ofrecer resultados útiles. Esto es, los constructores de tests deben ofrecer baremos apropiados que permitan interpretar las puntuaciones obtenidas o, al menos, claras indicaciones sobre cómo tipificar o baremar la prueba en distintos grupos normativos. Es evidente que si en un centro escolar se va a examinar a alumnos de primeros cursos de enseñanza primaria y el test sólo ofrece baremos para los últimos cursos, no puede considerarse útil en dicho centro al carecer de baremos de referencia. Relacionado con este hecho está el

tema de los errores de medida. Dichos errores se derivan del hecho mismo de considerar la puntuación de un sujeto en un test como estimación de un rasgo o característica. Dicha estimación implica la diferenciación entre puntuación verdadera y puntuación estimada de un sujeto en un test. Suele formularse, en términos de la teoría clásica, así:

$$X_{ij} = \text{Puntuación del test Puntuación verdadera} \pm \text{error}$$

Si los tests fueran absolutamente fiables (r= 1,00), no habría error en la estimación de un rasgo a través de ellos. En ciencias sociales y cuando medimos características humanas, sabemos que esto no es posible, de tal manera que aceptamos que al medir estamos siempre cometiendo un error de medida. Error por sobreestimación (se obtiene una puntuación superior a la verdadera) o por infraestimación (puntuación obtenida inferior a la puntuación verdadera para un sujeto). Esto es lógico, el sujeto puede acertar por azar - en un test de rendimiento, por ejemplo - más ítems de los que debería según sus conocimientos; o bien, fallar algunos ítems más de lo que debería (elementos mal formulados, situación de examen, etc.). Normalmente, y sobre bases de estadística inferencial, puede estimarse este error que se comete al medir - siempre que se considere aleatorio, no sistemático - con un determinado nivel de confianza o seguridad. Ello suele realizarse sobre la base del error típico de medida (desviación típica de la distribución muestral de los errores). Los manuales de los tests deberían ofrecer este dato, ya que ello permitiría a los usuarios de tests estimar los límites o intervalos de confianza de una puntuación verdadera en un test, conocida la puntuación obtenida en la prueba. La utilidad de estos intervalos de confianza de una puntuación verdadera, si el error de medida no es muy elevado, puede ser muy grande en modelos de enseñanza individualizada o personalizada. Veremos esto más despacio más adelante.

7.2. Instrumentos internos de medida del rendimiento

Al hablar de los instrumentos de medida interna, nos referimos a aquellos que son elaborados dentro de un determinado centro educativo y para uso del mismo. Son pruebas construidas por el profesor de una clase, por el equipo de profesores, por un departamento o, en general, por miembros de la entidad docente.

De hecho, la mayoría de los instrumentos utilizados en los procesos evaluativos, en el contexto de un centro escolar, son medidas internas, que deberán ser construidos y evaluados con sumo cuidado. Dichas medidas se construyen para evaluar conceptos cognitivos específicos implicados en las tareas escolares de instrucción, conceptos que no pueden ser medidos o evaluados a través de tests tipificados que, como mencionamos, están básicamente destinados a medir variables o aptitudes de tipo general. De aquí que las medidas internas no estén tipificadas, ni tendría sentido que lo estuvieran por no ir destinados a gran número de sujetos, sino a grupos de clase o lo más, a cursos completos

de un centro.

Para medir el rendimiento escolar o el aprovechamiento escolar, el aprendizaje en suma, pueden utilizarse recursos como entrevistas, observación, y pruebas de rendimiento. De los primeros tipos de recursos, ya hemos mencionado anteriormente sus características, posibilidades y limitaciones. Ahora vamos a analizar estas últimas, las pruebas de rendimiento, instrumentos que ofrecen ventajas que no poseen los demás recursos de recogida de información mencionados: pueden aplicarse en momentos puntuales adecuados, puede planificarse su estructura y su alcance, pueden aplicarse simultáneamente a grandes grupos, permiten una valoración o sistema de puntuación más objetivo, de aquí que sean el recurso más utilizado para la medición del aprovechamiento escolar. Nos referimos a las pruebas objetivas y de ensayo, sin duda alguna, el tipo de pruebas sobre las que recae mayoritariamente el peso de la medida del rendimiento escolar o académico.

Hoy por hoy, las mediciones conseguidas a partir de pruebas internas, constituyen el mayor peso de la materia prima que se utiliza para la evaluación; de aquí que la calidad, uniformidad y precisión de estos instrumentos deba garantizarse al máximo, dado que dan lugar a la toma de decisiones pedagógicas: la asignación de calificaciones, la adopción de alternativas de enseñanza o la modificación de los planes y programas de estudio. Como ya hemos señalado, por trabajar a base de mediciones, la evaluación está sujeta a errores procedentes de éstas. En el caso del aprovechamiento escolar, para minimizar estos errores, será preciso: trabajar con objetivos de aprendizaje bien especificados, técnicas e instrumentos bien diseñados y uniformes, y escalas de valoración, cuyos grados o niveles correspondan a condiciones de aprendizaje concretados y fácilmente discriminables.

Al abordar este capítulo, pues, nos vamos a detener fundamentalmente en el estudio de dos tipos básicos de pruebas: las pruebas objetivas y las pruebas de ensayo (también dedicaremos una breve referencia a las pruebas de resolución de problemas). En este enfoque, no debe perderse la visión general y verdadera de la evaluación del aprovechamiento o del rendimiento escolar, esto es, llega bastante más allá de la estricta medición del aprendizaje, siendo ésta sólo un antecedente de aquélla. Al objeto de no perder de vista el proceso completo ofrecemos la figura 7.1.

7.2.1. Referencia criterial y normativa

Esta dicotomía es una de las primeras cuestiones que deberá plantearse un profesor al decidir la redacción de un test o prueba de rendimiento con una finalidad específica. Como ya hemos mencionado, es la finalidad de la prueba la que debe decidir qué sistema de referencia elegir. Si se pretende evaluar si un sujeto posee las capacidades o conocimientos requeridos en un momento determinado de su desarrollo y pueden identificarse dichas capacidades, lo adecuado será elaborar un test criterial. Si, por el

contrario, el propósito es básicamente evaluar la situación del sujeto en relación con algún grupo normativo (específico), lo adecuado sería construir un test normativo.

Distintos autores como Pophan (1980), De la Orden (1982), Beggs y Lewis (1979), etc., coinciden al sostener que para la evaluación educativa, y más en concreto para la medición del nivel de rendimiento (logro, aprovechamiento), las pruebas tradicionales de refe rencia normativa son poco útiles. Por el contrario, la referencia criterial parece ofrecer una alternativa de evaluación del aprendizaje del alumno, más en consonancia con sus funciones pedagógicas.

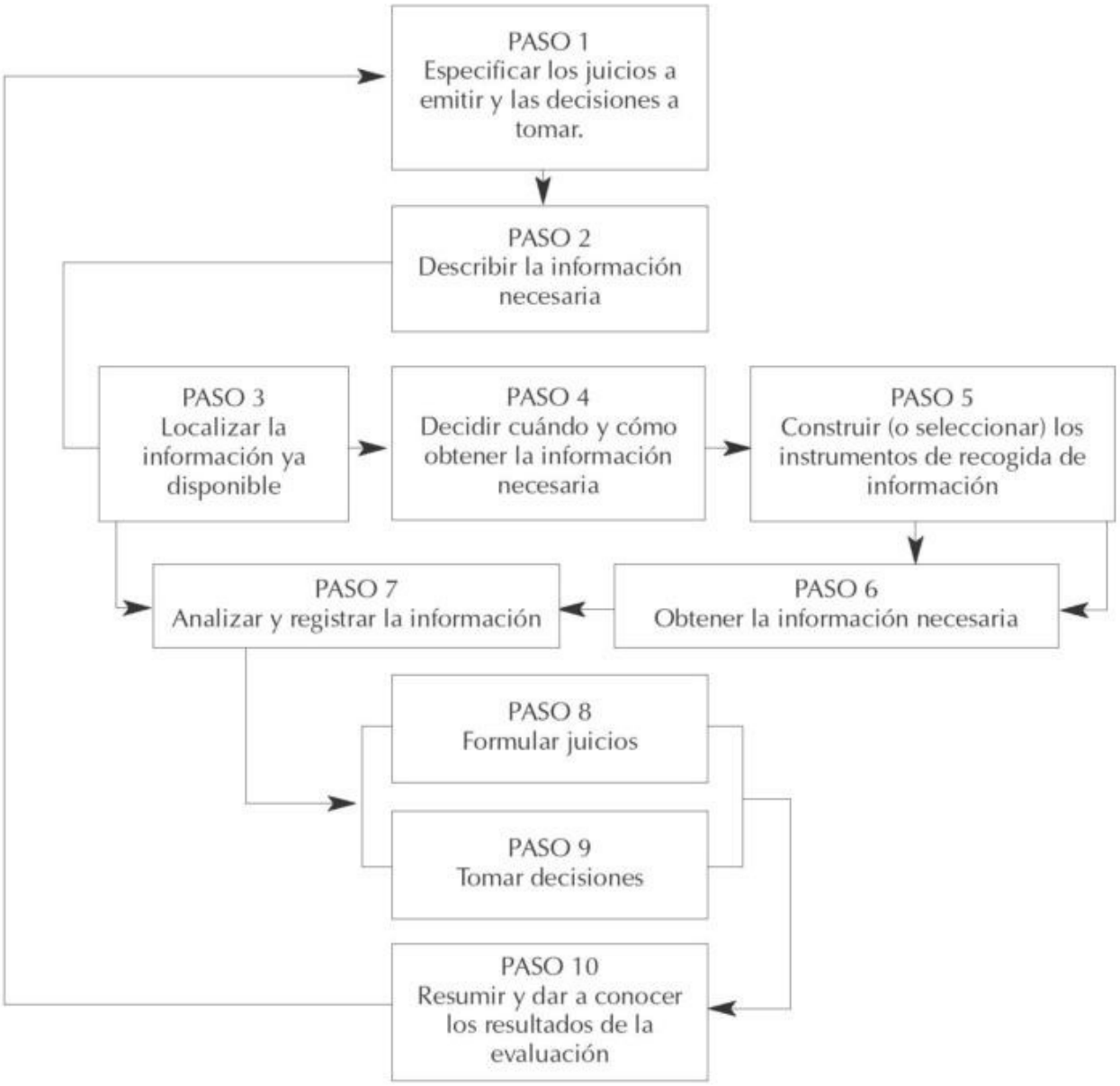


Figura 7.1. Pasos del proceso de evaluación (Tenbrick, 1981).

La esencia de la medición con referencia a un criterio, tal como la definió Glaser

(1963) y en palabras de Pophan (1980) reside en: a) La delimitación de un campo de conductas bien explicitado; b) La determinación de la actuación del individuo en relación con ese campo (pág. 151). Como ya mencionamos anteriormente, conviene no confundir "criterio" con: a) objetivo educativo o instructivo; b) nivel deseado de capacitación del alumno; o c) resul tado postinstruccional deseado. Al hablar de la actuación de una persona con referencia a un criterio, hablamos de referencia a una clase de conductas bien definidas (dominio).

Las ventajas que ofrece la referencia criterial, respecto a la normativa en la medición educativa han sido analizadas por muchos autores. Entre dichas ventajas destaca, sin duda, el hecho de no buscar dispersión en las respuestas, sino más bien describir al máximo y lo mejor posible la actuación de un sujeto respecto a un dominio o clase de conductas definidas. En este sentido, y como la investigación ya ha demostrado en alguna ocasión, las pruebas referidas a un criterio serán generalmente más sensibles para detectar efectos de la enseñanza, lo que sin duda es de un valor muy apreciable para el profesor o evaluador.

De lo anterior se desprende que a la hora de valorar pruebas referidas a criterio, lo más importante es que dichas pruebas tengan validez de contenido, esto es, congruencia entre las conductas implícitas en los elementos de la prueba y las definidas en las metas y objetivos del proceso de enseñanza-aprendizaje. De aquí que el profesor o constructor de estas pruebas deba determinar si los elementos que pretende incluir tienen valor evaluativo, miden lo que pretenden medir. Para conseguir esta congruencia (De la Orden, 1982) "es necesario seguir escrupulosamente una serie de pasos especificados en algoritmos o series muy precisas de instrucciones para la construcción de las pruebas, siendo el más importante de todos, como es obvio, la transformación de la intención didáctica de los objetivos en preguntas para las pruebas, esto es, el establecimiento de una relación directa entre cada ítem y el objetivo cuyo logro intenta medir" (pág. 80).

La fiabilidad de un test referido a criterios tiene poca importancia; tengamos en cuenta que al construirse no se busca, por ejemplo, que haya correlación entre dos aplicaciones sucesivas de la prueba (test-retest), si un sujeto fracasa en ciertos elementos del test el profesor dirigirá su atención docente a los conceptos medidos en dichos elementos, no esperando ni deseando que, en una segunda medición, el sujeto puntúe de la misma forma. Tampoco se espera consistencia interna (otra forma de estimar la fiabilidad de un test) entre los elementos de un test criterial, dado que cada elemento del mismo se define como criterio de una capacidad o dominio a evaluar, no siendo esperable necesariamente intercorrelación entre los elementos. Por otro lado, tampoco se desea especialmente que un test referido a criterio discrimine entre sujetos, esto es, un elemento puede ser válido aunque todos los sujetos contesten correctamente o todos incorrectamente. Si el elemento tiene validez de contenido (mide lo que desea medir) y todos los alumnos, por ejemplo, lo contestan correctamente, el elemento sigue siendo bueno y valioso para un test criterial, aunque no discrimine (por el contrario, en un test referido a normas, dicho

elemento sería considerado deficiente). De aquí se deriva la recomendación de que: "el profesor se asegure sólo de que cada elemento de un test criterial tenga validez de contenido...; si la tiene, la fiabilidad será una consecuencia implícita" (Beggs y Lewis, 1979, pág. 170).

Existen diversos procedimientos para generar elementos de una prueba de referencia criterial, a partir de las conductas expresadas por los objetivos. Siguiendo a De la Orden (1982, págs. 80-82) cabe destacar los siguientes:

a)Derivación directa del ítem a partir del objetivo. El problema de este procedimiento está relacionado con el alto grado de subjetividad en la interpretación del objeti vo, especialmente si su formulación, como es frecuente, no cumple los requisitos de precisión y especifidad al más alto nivel. Analizar el significado conductual del verbo principal de la proposición que expresa el objetivo, suele resultar insuficiente, dada la dificultad de determinar la conducta específica implicada en el verbo y, en muchos casos, el nivel de ambigüedad y polisemia de los verbos. Una condición que se exige, básicamente para evitar este problema de subjetivismo, es contrastar la denominada "bondad de ajuste" del ítem con las prescripciones de una buena taxonomía de objetivos (la de Bloom, por ejemplo), lo que ayudará al evaluador a decidir con más precisión lo que significa el verbo del objetivo, a fin de redactar elementos que requieran del alumno la misma conducta. Otras condiciones son: verificar la "validez de contenido" del elemento, asegurar que los elementos cubran al máximo el dominio o campo conductual definido en el objetivo y que el contenido de los elementos sea relevante y significativo.

b)Derivación mediante el previo desarrollo de un esquema básico o modelo general de elementos que defina con claridad el dominio de realizaciones o conductas que la prueba debe medir.

c)Derivación a partir de reglas de transformación de textos, relacionados con los principios de transformación lingüística, definiendo así específicamente los tipos de elementos que se han de generar.

En principio y en la construcción de programas de examen o evaluación del rendimiento de los centros educativos, caben tests construidos con ambos sistemas de referencia (criterial y normativa), siendo probablemente el más útil el sistema de referencia criterial, mientras que el de referencia normativa debiera restringirse más a casos en que se pretendiera aplicar instrumentos en varios centros. De todos modos, ambos tipos de tests muestran ventajas y limitaciones. El profesor, departamento o centro educativo es el que deberá decidir cuál de los sistemas de referencia es más útil, en cada situación, para mejorar el aprendizaje de los alumnos.

7.2.2. La prueba objetiva

En este apartado vamos a dedicar nuestra atención a las pruebas objetivas, caracterizadas por estar construidas a base de ítems, elementos o reactivos, cuya respuesta no deja lugar a dudas respecto a su corrección o incorrección. En ellas, el estudiante trabaja sobre una situación estructurada a la que no aporta más que respuestas concretas. Puede utilizarse con fines diagnósticos, formativos o sumativos. Puede elaborarse como prueba tipificada elaborada con referencia a normas y con todos los requisitos técnicos: fiabilidad, validez, baremos, etc.); o bien, como una prueba informal, elaborada para propósitos específicos de un profesor relativos al curso que imparte y a los alumnos que participan en él, etc. También puede pensarse en la elaboración de ítems o elementos para una prueba criterial, en cuyo caso cada ítem será construido de acuerdo al dominio o campo de conductas que pretenda medir. En estos casos anteriores y, dadas sus diferentes concepciones, habrá que tener en cuenta las mismas variables que cada tipo de prueba impone.

Excede los propósitos de este trabajo exponer por separado las distintas fórmulas posibles de instrumentos de medición, de aquí que nos referiremos casi exclusivamente a las pruebas objetivas informales, exponiendo comentarios orientativos, según sirvan para distintos propósitos: diagnósticos, formativos o sumativos.

Pretendemos, sobre todo, ofrecer al profesor apoyo técnico para que construya con las mayores garantías posibles de rigor y objetividad, pero sin entrar en el rigor de elaboración que exigirían pruebas objetivas tipificadas, que por otro lado no serían demasiado útiles, dado el largo tiempo que precisan para su elaboración, los controles técnicos que exigen y la limitada aplicabilidad que tendrían, pues en la mayoría de las ocasiones las condiciones de enseñanza-aprendizaje cotidianas son poco estables, con contenidos cambiantes, para grupos de alumnos en diferentes situaciones educativas y bajo presiones administrativas que deforman el sentido real de la evaluación.

En todo caso, las características de las pruebas tipificadas y de las pruebas informales son semejantes, sólo empiezan a diferenciarse a partir de las distintas situaciones educativas de aplicación o bien, por la exigencia o no exigencia del cumplimiento de requisitos técnicos (fiabilidad, validez, índices estadísticos de los ítems, etc.). No hay importantes diferencias relativas al cómo fueron o deben ser construidos los ítems o reactivos.

A) Planificación de una prueba objetiva

a)Propósito y contenidos de la prueba

Las primeras preguntas que se han de responder cuando queremos construir una prueba objetiva serán: ¿para qué se utilizará la prueba? ¿Qué aprendizaje o área de contenidos exploraremos con ella?

A la primera cuestión habrá que responder diferenciando entre propósitos diagnósticos, formativos o sumativos. Diagnósticos, esto es, exploración de aquellos aprendizajes que el alumno debiera poseer antes de acometer un nuevo curso o una nueva área de aprendizaje. Formativos, es decir, exploración de los aprendizajes derivados de las experiencias educativas ya efectuadas, intentando valorar aquellos conocimientos y aprendizajes que esperábamos conseguir en dichas experiencias. Sumativos, o exploratorios del aprendizaje derivado de todo un curso o de un campo amplio y unitario de conocimientos y aprendizajes, intentando incluir una muestra representativa, suficiente y equilibrada de todos los contenidos objetivos definidos para el curso o campo de aprendizaje evaluado.

Sin embargo, y sea cual sea el propósito de la prueba, deberá responderse a la segunda cuestión, a través de una representación esquemática de los temas, contenidos y objetivos que la prueba deberá incluir.

b)Importancia relativa de los contenidos

En dicho esquema, aparecerá ponderada la importancia relativa de cada contenido dentro del total de cuestiones o contenidos a evaluar (una fórmula sencilla es asignar porcentajes a cada tema, de manera que en total sumen el 100 por 100). Es preferible que dicha asignación de porcentajes se realice mediante el acuerdo de varios profesores para evitar subjetivismos o influencias personales. Como ejemplo presentamos el siguiente cuadro.

CUADRO 7.7. Importancia relativa de cuestiones, contenidos o temas dentro de un área de contenido a evaluar (Carreño, 1977)

Asignatura: "Metodología de la Ciencia"
Capítulo: "Hipótesis científica"

Temas	*Pesos 0% de importancia relativa*
1. Reconocimiento de los pasos del método científico	6
2. Caracterización de problemas científicos	10
3. Reglas para el correcto planteamiento de problemas	10
4. Localización de problemas científicos	6
5. Definición de hipótesis científica	10
6. Importancia de las hipótesis en la investigación	10
7. Condiciones para formular correctamente hipótesis	20
8. Contrastabilidad de la hipótesis	10
9. Distinción entre contrastabilidad formal y empírica	8
10. Las técnicas de contrastación	10
	Σ = 100%

En el cuadro anterior, la asignación de pesos o porcentajes debe indicar importancia relativa de unos temas respecto de otros. Está claro que el tema 7 es el más importante y valioso, tanto como el doble que cualquiera de los demás. Los temas 1, 4 y 8 son menos valiosos que el resto de ellos, el 1 y el 4 menos importantes que el 8.

c)Número de elementos o reactivos a incluir

Ponderados los contenidos a evaluar será necesario determinar el número de ítems a incluir en la prueba. Suele aconsejarse no incluir menos de 50 ítems si la prueba es sumativa y no menos de 20 si la prueba es formativa. Incluir menos ítems es dejar demasiado margen al azar en la medición del área de contenido, lo que restará certidumbre al juicio evaluativo. Suponiendo, por ejemplo, que queremos incluir 80 ítems y continuando con el ejemplo de Carreño (1977), el cuadro 7.8 refleja el reparto de los 80 ítems teniendo en cuenta la importancia relativa de cada tema.

CUADRO 7.8. Reparto ponderado de ítems en función de la importancia relativa de cada tema

Tema	*Peso relativo*	*N.° de ítems*	*N.° de ítems ajustado*
1	6	4,8	5
2	10	8	8
3	10	8	8
4	6	4,8	5
5	10	8	8
6	10	8	8
7	20	16	16
8	10	8	8
9	8	6,4	6
10	10	8	8
	Σ = 100	Σ = 80	Σ = 80

d)Número de elementos de cada tema correspondiente a cada nivel de la taxonomía de objetivos educativos

El siguiente paso, una vez determinados cuántos ítems habrá que construir para cada subárea de contenido (que venimos llamando tema), es la determinación de qué objetivos - dentro de una taxonomía de objetivos - deberán medirse a través de los ítems (elementos o reactivos) correspondientes a cada tema.

Ello suele realizarse en dos pasos sucesivos:

1. Construcción de una tabla de especificaciones (objetivos-contenidos).

2. Construcción de una tabla de correspondencias entre objetivos-contenidos-número de ítems para cada unidad temática o tema.

Siguiendo el ejemplo anterior y utilizando como base la taxonomía de Bloom, la tabla de especificaciones podría ser la que muestra el cuadro 7.9.

CUADRO 7.9. Especificaciones que relacionan objetivos y contenidos

Nivel cognitivo de objetivos	*Temas (contenidos)*									
	1	2	3	4	5	6	7	8	9	10
1. Conocimiento	X	X	X	X	X	X	X	X	X	X
2. Comprensión		X	X	X	X	X	X	X	X	X
3. Aplicación			X	X	X	X	X	X	X	X
4. Análisis							X	X		X
5. Síntesis							X	X		X
6. Evaluación								X		X

El cuadro 7.9 muestra, a través de 35 cruces (X) los objetivos que deben conseguirse en referencia a cada contenido o tema. Por ejemplo, con relación al tema 4, deberán perseguirse objetivos correspondientes a los niveles cognitivos de CONOCIMIENTO, COMPRENSIÓN y APLICACIÓN; para el tema 7 deberán conseguirse - además de los niveles anteriores - los niveles cognitivos de ANÁLISIS y SÍNTESIS, etcétera.

No obstante, la información de este cuadro habrá que completarla determinando si se construirán ítems o reactivos para cada cruz (x); en caso positivo, cuántos para cada cruz (X) y en caso negativo, qué cruces de objetivo-contenido tendrán reflejo directo en cada ítem. Obviamente, responder a esta cuestión exige contemplar varios aspectos:

1. ¿Cuántos ítems tendrá la prueba? (siguiendo nuestro ejemplo 80).

2. ¿Cuántos ítems corresponden a cada tema? (en nuestro caso dicha información se encuentra en el cuadro 7.8).

3. Dentro de cada tema, ¿hay suficientes ítems para cubrir todos los cruces

objetivoscontenido? (en nuestro ejemplo, sí):

a)En caso positivo, construir al menos un ítem para cada cruz, los ítems sobrantes deberán asignarse a los niveles cognitivos (objetivos) que se consideren de mayor relevancia relativa, o bien, si ello no es posible o no está claro, se distribuirán al azar entre los cruces objetivo-contenido señalados para ese tema.

b)En caso negativo, deberán asignarse ítems a los niveles cognitivos (objetivos) que se consideren de mayor relevancia relativa con relación al tema o área de contenido particular. Si ello no es posible o no está claro, se distribuirán al azar entre los cruces objetivo-contenido señalados para ese tema.

Tomadas las decisiones correspondientes con relación a nuestro ejemplo, construimos un último cuadro de correspondencias objetivos-contenidos. En las casillas o cruces aparece el número de ítems asignado a cada conducta a medir (cuadro 7.10):

CUADRO 7.10. Correspondencia contenidos-objetivos-número de ítems asignados a cada conducta a evaluar.

Nivel cognitivo (objetivos)	*Temas (contenido)*										*Total de ítems por nivel cognitivo*
	1	2	3	4	5	6	7	8	9	10	
1. Conocimiento	X5	X4	X3	X2	X3	X2	X3	X2	X2	X2	28
2. Comprensión		X4	X3	X2	X2	X3	X4	X2	X2	X2	24
3. Aplicación			X2	X1	X3	X3	X3	X1	X2	X1	16
4. Análisis							X3	X1		X1	5
5. Síntesis							X3	X1		X1	5
6. Evaluación								X1		X1	2
N.° de ítems asignados a cada tema	5	8	8	5	8	8	16	8	6	8	80

Siguiendo estos criterios se consigue que la prueba se encuentre equilibrada y sea un muestreo suficiente, relevante y representativo del área de conductas a evaluar. La siguiente tarea será construir ítems que midan adecuadamente las conductas definidas en la tabla anterior.

B) Características generales que deben reunir los elementos o ítems de una prueba objetiva

De modo orientativo, enunciamos aquí una serie de características que deben reunir buenos elementos de una prueba objetiva, sea cual sea el tipo de elemento de que se

trate:

1. Corresponde en contenido al objetivo de aprendizaje para el que fue elaborado.

2. Corresponde en tipo y procedimiento de solución al aprendizaje o aprendizajes del objetivo, en cuanto ala expresión conductual en que su logro se manifiesta.

3. No debe sobrepasar el nivel de exigencia, complejidad y extensión con que fue tratado el tema, al ser aprendido por el alumno. Su grado de dificultad es acorde con el criterio de logro del aprendizaje que se espera.

4. No debe formularse en términos o contextos nuevos para el estudiante. Si se usan términos no familiares, no se están probando los conocimientos señalados, sino que se evalúa vocabulario desconocido.

5. Debe formularse en términos claros, precisos y sin ambigüedades (no debe exigir interpretación gramatical). En este sentido, está exento de "trampas" o "segundas intenciones".

6. Utiliza vocabulario y formas sintácticas que los alumnos pueden comprender sin problema de lectura. La aptitud lectora no debe influir en los resultados de la prueba.

7. No debe formularse en los términos con que aparece la noción en el libro de consulta o textos; de este modo se estimularía el aprendizaje puramente memorístico de frases y párrafos.

8. No debe formularse en términos que den claves para responder a otro u otros ítems.

9. Es "independiente" de otros elementos. No es deseable incluir cuestiones que sólo pueden responderse correctamente si se ha acertado en otra u otras anteriores.

10. Exige una respuesta que permite al examinador valorar si el objetivo de referencia o los aprendizajes correspondientes se han logrado o no.

C) Tipos de ítems

Una clasificación que suele realizarse de las pruebas objetivas está en relación con el tipo de ítems que se utilizan en ellas: de respuesta abierta (de elaboración de respuesta) o de respuesta cerrada (de reconocimiento o elección). En este apartado nos referiremos fundamentalmente a los primeros, y dentro de los segundos a los ítems de verdadero/falso. A los ítems de opción múltiple (respuesta cerrada), por su especial relevancia, les dedicaremos un apartado particular.

a)De respuesta breve y complementación

Son aquellos que exigen una palabra o un símbolo para su contestación. Son muy elementales y miden básicamente memorización.

Los datos son muy concretos, por lo que una prueba objetiva no debe estar constituida únicamente por ítems de este tipo:

-Ejemplos:

•¿Cómo se llama el inventor de la imprenta? (Guttemberg).

•El inventor del gramófono fue... (T.A.Edison).

-Ventajas: fácil construcción. Dificulta la adivinación del alumno en caso de no conocer la respuesta.

-Inconvenientes: inadecuación para medir productos complejos de aprendizaje. Dificultad de calificación en ocasiones (por ejemplo, en la pregunta: "¿Dónde nació Napoleón?", ¿qué se pregunta: nación, continente, ciudad, región, etc.?).

b)De respuesta alternativa (sí/no; verdadero/falso)

Limitan la respuesta a dos alternativas. Miden capacidad de identificar conceptos muy precisos definidos de forma similar, o de discriminar lo correcto de lo incorrecto.

-Ejemplos:

•La tierra es un planeta F O VO.

•La tierra gira alrededor de la luna F O VO.

•La tierra gira alrededor de: la luna (/el sol O.

-Ventajas: fáciles de construir. Pueden permitir discriminaciones finas del dominio de conceptos.

Para eliminar el efecto del azar en la respuesta, se corrige utilizando la fórmula:

Puntuación total en la prueba = Aciertos – Errores

Previamente los examinados deben ser, obviamente, avisados de ello.

Muchos autores de tests procuran evitar la utilización de este tipo de ítem, al considerarlo poco útil, pues al construirlos suele tenderse hacia la excesiva facilidad (dar pistas para acertar) o bien hacia la excesiva dificultad (incluir varios elementos a considerar para decidir sobre la veracidad o falsedad de la aseveración).

c)De ordenamiento o jerarquización

Se caracterizan por ofrecer a los sujetos a evaluar listas de datos o elementos que deben ordenar o jerarquizar de acuerdo con algún criterio que se indica (cronológico, lógico, evolutivo, etc.).

Permiten medir aprendizajes de distintos niveles cognitivos (desde conocimiento hasta aplicación, normalmente). Pueden presentar problemas a la hora de ser puntuados.

-Ejemplo:

Ordenar de mayor a menor dureza los siguientes minerales: topacio, cuarzo, feldespato, diamante y hulla.

-Ejemplo:

Ordenar jerárquicamente los niveles de aprendizaje propuestos por Bloom en el dominio cognoscitivo empezando por el más elemental: comprensión, evaluación, aplicación, síntesis, conocimiento y análisis.

d)De asociación, emparejamiento o correspondencia

En ellos se pide al sujeto que establezca relaciones entre elementos de dos series o grupos. Requiere instrucciones muy claras, pudiendo, en ocasiones, resultar que varios elementos de una serie correspondan a uno de la otra.

-Ejemplo:

Anotar en los paréntesis de la columna A (países) los números de los elementos de la columna B (capitales) que correspondan:

Columna A	*Columna B*
( ) Japón	(1) Pekín
( ) India	(2) Tel Aviv
( ) China	(3) Katmandú
( ) Nepal	(4) Nueva Delhi
( ) Israel	(5) Tokio
	(6) Jerusalén

Normalmente, debe haber más posibilidades en la columna de la derecha para evitar respuestas por eliminación.

Estos ítems pueden utilizarse para comprobar el aprendizaje de relaciones entre, por ejemplo: teoría/autores, fórmulas/fenómenos, fechas/obras, lugares/personajes, etc. Es un ítem que puede adoptar múltiples formas y presentaciones; útil para una gran variedad de niveles de objetivos de aprendizaje, especialmente en aquellos que requieran: vincular, relacionar, clasificar, aplicar leyes o principios, inferir, etc.

e)De identificación o localización

Se caracterizan por contener instrucciones referidas a conductas que han de darse sobre esquemas, gráficos o cuadros (mapas, diagramas, cuadros sinópticos, modelos, fotos, ilustraciones, etc.). Suelen requerir localizar o reconocer elementos en un contexto visual.

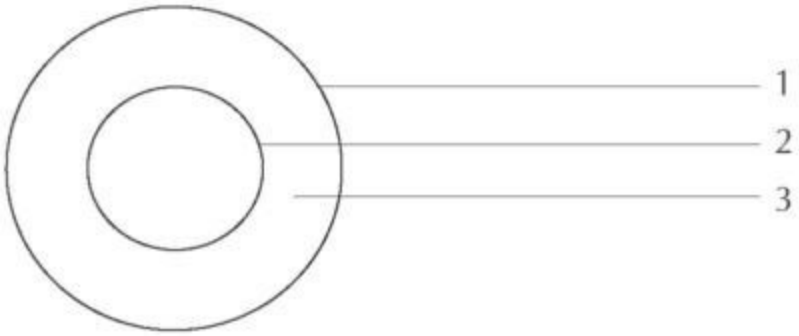
Miden la capacidad de relacionar y bien utilizados pueden ser idóneos para comprobar o medir la consecución de aprendizajes complejos o de alto nivel (identificar ondas en un electrocardiograma, diferenciar zonas urbanas, desérticas y selváticas sobre un mapa, etc.).

-Ejemplo:

Identificar los elementos de la célula.

-O bien:

Identifica en esta célula, el núcleo(), el citoplasma() y la membrana().



Suele utilizarse mucho en áreas de conocimiento como geografía o ciencias naturales.

Debe cuidarse mucho que el material gráfico presentado para la medición de aprendizaje sea similar o igual al empleado en las experiencias de aprendizaje, a no ser que, por sí mismo, se busque algún tipo de transferencia de aprendizaje a través de material novedoso.

Puede también pensarse en la utilización de diapositivas (por ejemplo, historia del arte) o material auditivo (reconocimiento de sonidos, piezas musicales, instrumentos, etc.).

D) Elementos de lección u opción múltiple

Este tipo de ítem suele reconocerse como el más útil y más ampliamente aplicable dentro de una prueba objetiva. Puede medir gran variedad de productos de aprendizajes, simples o complejos y en distintas áreas de conocimiento.

A través de ellos pueden alcanzarse más fácilmente niveles más altos dentro de una jerarquía (taxonomía) de objetivos de aprendizaje.

Dado que influye el azar en la respuesta, suelen corregirse eliminando de la puntuación total dicha influencia a través de la siguiente fórmula:

$$\text{Puntuación total} = A - \frac{E}{K-1}$$

donde A = n.º de aciertos.
E = n.º de errores.
K = n.º de alternativas.

a)Características

Se definen por constar de un enunciado o problema inicial y una lista de soluciones

(opciones) sugeridas. El enunciado puede plantearse en forma de pregunta directa o bien en forma de aseveración inconclusa. Algunos autores denominan vástago o tronco del ítem a esta parte. La lista de soluciones que se sugiere puede incluir palabras, números, símbolos o frases, y se denominan alternativas de respuesta.

-Ejemplo:

Vástago de pregunta directa: ¿Cuál de las siguientes es la capital de Islandia?

a) Copenhague	c) Oslo
b) Reykjavik	d) Helsinki

-Ejemplo:

Vástago de aseveración inconclusa: La capital de Islandia es...

a) Copenhague	c) Oslo
b) Reykjavik	d) Helsinki

En los ejemplos anteriores sólo hay una respuesta correcta, las demás son erróneas (suele denominarse distractores a las respuestas erróneas). Pueden existir otros conocimientos en que puede no haber términos tan precisos de respuesta, por ello pueden formularse ítems que exigen elegir la mejor respuesta o solución a una cuestión:

-Ejemplo:

¿A cuál de los siguientes factores se le suele dar mayor consideración al seleccionar una ciudad como capital de un estado?

a) Clima	c) Situación geográfica
b) Carreteras	d) Número de habitantes

El tipo de ítem de respuesta correcta es especialmente aconsejable cuando se trata de evaluar terminología o conocimientos específicos:

-Ejemplo:

¿Si un test mide lo que dice medir, se dice que tiene...?

a) Fiabilidad	c) Validez
b) Utilidad	d) Relevancia

-Ejemplo:

Los tests construidos con referencia a norma se denominan también...

a) Internos	c) Tipificados
b) Criteriales	d) Sumativos

b)Posibilidades, ventajas y limitaciones

Este tipo de ítems, como ya se ha dicho, permite medir productos complejos de aprendizaje (gran flexibilidad).

Eliminan la ambigüedad propia de los ítems anteriores.

Su aplicación requiere menor tiempo, pudiéndose aumentar el número de ítems y abarcar así mayor parte del dominio o universo de conocimientos a evaluar.

Presentan gran eficacia en la evaluación de conocimientos terminológicos y específicos. Su aplicabilidad es muy amplia, pudiéndose utilizar para la medición de productos en los niveles de comprensión y aplicación, habilidades para interpretar relaciones causaefecto, para justificar métodos y procedimientos, para enjuiciar hechos y situaciones, diferencias, semejanzas, ordenación y jerarquía, identificación de errores, síntesis de principios, etc.

Su corrección es muy rápida, incluso puede mecanizarse.

Permite corregir el efecto del azar en la respuesta.

A pesar de su superioridad sobre otros tipos de ítems, presentan también algunas limitaciones, entre las que destacan:

-No es aconsejable utilizarlos para medir la resolución de habilidades para resolver problemas en matemáticas o ciencias. Tampoco es apropiado su uso para medir la capacidad para organizar y presentar ideas.

-Son más difíciles de construir que algunos de los tipos de ítems anteriores.

-En ocasiones no es fácil encontrar buenos "distractores" (alternativas erróneas pero plausibles).

c) Sugerencias para construir ítems de elección múltiple

Los ítems de opción múltiple no son fáciles de estructurar. La habilidad y la experiencia en la redacción de este tipo de elementos son muy importantes. De hecho, la elaboración de buenas pruebas objetivas lleva su tiempo.

Ofrecemos a continuación una serie de sugerencias o consideraciones a tener en cuenta al construir este tipo de ítem:

1. El vástago o tronco del elemento debe por sí mismo tener significado, presentar una forma definida.

-Deficiente:

El País Vasco:

a)Es una tierra árida y plana.

b)No tiene industria.

c)Tiene más habitantes que Cataluña.

d)Tiene costa en el mar Cantábrico.

-Mejor:

El País Vasco tiene sus costas en el Mar:

a)Mediterráneo.

b)Atlántico.

c)Cantábrico.

d)Del Norte.

2. El vástago del elemento debe incluir tanto como sea posible del contenido del elemento y estar exento de material que no sea pertinente.

-Deficiente:

La mayor parte de Sudamérica la colonizaron individuos que procedían de España, ¿cómo se explica el gran número de españoles que fueron a colonizar esas tierras?

a)Les gustaban las aventuras.

b)Buscaban riquezas.

C)Querían menos impuestos.

d)Iban en pos de la libertad religiosa.

-Mejor:

Los españoles colonizaron la mayor parte de Sudamérica. La razón fundamental era que:

a)Les gustaban las aventuras.

b)Buscaban riquezas.

c)Querían menos impuestos.

d)Iban en pos de la libertad religiosa.

3. Procurar evitar, en lo posible, la formulación negativa del vástago del elemento. Sólo si los productos de aprendizaje lo requieren estaría justificada.

-Deficiente:

¿Cuál de los siguientes países no está al norte del Ecuador?

a)España.

b)Turquía.

c)México.

d)Angola.

-Mejor:

¿Cuál de los siguientes países esta al sur del Ecuador?

a)España.

b)Turquía.

c)México.

d)Angola.

4. Todas las alternativas deben ser gramaticalmente congruentes con el vástago del elemento.

-Deficiente:

Un alto índice de fiabilidad de un test indica:

a)Este test es muy válido.

b)Alto nivel de precisión en la medida.

c)No puede hallarse sin validez.

d)Estabilidad de las medidas del criterio.

-Mejor:

Un alto índice de fiabilidad de un test indica:

a)Validez de contenido.

b)Consistencia de los baremos del test.

c)Alto nivel de precisión en la medida.

d)Estabilidad de las medidas del criterio.

5. Un elemento o ítem debe contener únicamente una contestación correcta o una que sea evidente y claramente la mejor.

6. Los elementos usados para medir nivel cognitivo de "comprensión" deben contener "cierta" novedad, pero es preciso cuidarse de demasiadas novedades.

Las situaciones deben ser nuevas para los alumnos, pero no excesivamente desligadas de los ejemplos ilustrativos usados al explicar los conceptos en la clase. Un

ejemplo adecuado puede ser el mencionado por Bloom (1956) al medir la noción sobre el hecho de que "los materiales en el centro de la tierra están muy calientes". No deberían introducirse - si no se hizo en la explicación - en la prueba términos como "ígneo" o "fusión".

7. Todos los "distractores" o alternativas incorrectas de respuesta deben ser plausibles (lógicos):

-Deficiente:

¿Quién descubrió el Polo Norte?

a)Cristóbal Colón.

b)Fernando de Magallanes.

c)Robert Peary.

d)Marco Polo.

-Mejor:

¿Quién descubrió el Polo Norte?

a)Roald Amundsen.

b)Richard Byrd.

c)Robert Peary.

d)Robert Scott.

8. Deben evitarse asociaciones verbales entre el vástago y la respuesta correcta.

9. La longitud relativa de las alternativas no debe orientar o suministrar información o indicio alguno de la respuesta correcta.

-Deficiente:

La validez de constructo de un instrumento de medida permite:

a)Conocer si el test tiene ítems válidos.

b)Caracterizar el rasgo definido por la teoría base para la elaboración del instrumento.

c)Conocer si el criterio de test mide lo mismo.

d)Afirmar que el test mide exactamente el rasgo.

-Mejor:

La validez de constructo de un instrumento de medida permite:

a)Conocer si el test está formado por ítems que hacen referencia directa al rasgo a medir.

b)Caracterizar el rasgo definido por la teoría base para la elaboración del instrumento.

c)Determinar si el criterio de validación tiende a medir lo mismo que el test.

d)Afirmar que el test mide exactamente aquel rasgo que se propuso medir en un principio.

10.Debe procurarse evitar una pauta regular de respuestas correctas. Por ejemplo, en páginas 1, 3, 5, 7 y 9, la respuesta correcta se encuentra siempre en las posiciones 1.a, 2.a, 3.a, 4.a y 5.a, respectivamente.

11.Procurar usar lo menos posible alternativas especiales de respuesta del tipo: "ninguna de las anteriores", "a y c", "todas las anteriores", etc.

12.Evitar la construcción de ítems o elementos interdependientes, esto es, elementos que sólo pueden responderse acertadamente si se ha respondido bien a algún elemento anterior.

13.En general, todo elemento de opción múltiple debe consistir en una afirmación o en una idea, y no en una simple palabra. En tales casos, sería mejor utilizar, por ejemplo, elementos de respuesta breve o de verdadero-falso.

14.En el sentido de la anterior recomendación, conviene evitar el uso de elementos de opción múltiple allí donde otros tipos de elementos o ítems sean más apropiados.

De un modo orientativo y para facilitar la construcción de ítems de opción múltiple, presentamos a continuación el cuadro 7.11.

E) Integración de elementos al elaborar una prueba

De los apartados anteriores puede deducirse que construir pruebas objetivas no es tarea rápida y fácil. Elaborar ítems es hasta cierto punto sencillo, especialmente siguiendo las

sugerencias y normas anteriores; ahora bien, construir buenos ítems en relación con los objetivos de un curso es algo dificil de lograr por la cantidad de factores que participan en el proceso, condicionando así la calidad del resultado.

Como afirma Carreño (1977), "lo que hace bueno, útil y valioso a un reactivo o ítem es una combinación de su correcta construcción, su correspondencia con los logros que se busca medir y su integración equilibrada al resto de reactivos incluidos en la prueba" (pág. 60).

Así pues, cuanto más nos acerquemos al cumplimiento de las características enunciadas en los puntos B y D de este apartado, mejor construidos estarán los ítems. No debe olvidarse que, para construir buenos ítems, además de lo anterior, es necesario: dominio de la materia a evaluar, bastante entrenamiento y experiencia. No debemos pues desanimarnos si al principio cuesta mucho o bien no conseguimos elaborar buenos elementos; con sucesivas revisiones irán depurándose y acercándose a las formas preferibles.

A la hora de integrar los distintos elementos construidos, en una prueba única, pueden seguirse varios criterios. Dichos criterios han sido criticados y defendidos, por ello nos limi tamos aquí, siguiendo básicamente a Gronlund (1973) y a Carreño (1977), a plasmar una serie de sugerencias generales razonables, sabiendo que en casos particulares puede estar justificado no seguir algunas de ellas al pie de la letra.

CUADRO 7.11. Tipos de alternativas representativas a usar al medir cada uno de los niveles del ámbito cognoscitivo (Tenbrick, 1981)

Nivel cognitivo de aprendizaje	*Tipos de alternativas a presentar*	
1. Nivel de conocimiento	Definiciones Hechos Acontecimientos Nombres	Fechas Propiedades Estilos Tácticas
2. Nivel de comprensión	Ilustraciones Reenunciados	Interpretaciones Reordenaciones
3. Nivel de aplicación	Aplicación Generalizaciones	Reestructuración Extensión
4. Nivel de análisis	Análisis Elementos Suposiciones Relaciones Estructuras	Partes Temas Órdenes Modelos
5. Nivel de síntesis	Conceptos Taxonomías Esquemas Descubrimientos	 Planes Comunicaciones Invenciones
6. Nivel de evaluación	Inicios Fallos Puntos fuertes	Realizaciones Productos Puntos débiles

a) Tipos de elementos

Sin forzar la planificación inicial de la prueba, ni las circunstancias naturales de la conducta que cada objetivo exija, suele considerarse adecuado incluir en cada prueba varios tipos de ítems (de dos a cuatro) pero no demasiados. Ello, aunque suele alargar el tiempo necesario de corrección, supone ofrecer al alumno distintas variantes y posibilidades de operación mental y, por tanto, distintas oportunidades de que demuestre sus logros, dado que cada tipo de ítems implica distinto nivel u operación. Por otra parte, se facilita la expresión adecuada de los aprendizajes que se pretende medir, ya que no hay ningún tipo particular de ítem que permita medir -al mayor nivel de eficiencia posible - todos los distintos objetivos de niveles cognitivos diferentes.

b) Cantidad total de elementos

A esta cuestión ya respondimos anteriormente al pensar en la planificación del test o

prueba. Resumiendo, depende de los objetivos de la prueba. En general, tantos como sean precisos para explorar el logro de los objetivos programados respecto a un área de contenido particular. Al menos, uno por cada conducta (por cada aprendizaje contenido en cada objetivo).

También es importante valorar: tiempo de examen, tiempo de que se dispone para elaborarla y calificarla, tipo de examen (diagnóstico, formativo o sumativo), categoría de examen (parcial o final), etc. No hay normas fijas o parámetros para valorar la corrección o incorrección de una prueba, sin considerar los parámetros o circunstancias anteriores.

c) Organización de los elementos dentro de la prueba

Si hay elementos de varios tipos, conviene agruparlos según el tipo a que correspondan, lo cual facilita la inclusión de instrucciones y hace más fácil la corrección de la prueba.

Dentro de cada tipo de ítems, conviene ordenarlos de menor a mayor dificultad, con el fin de motivar al alumno en la resolución de la prueba. No se trata de "cazar" al alumno, sino de que la prueba sea capaz de detectar el "nivel" de aprendizaje al que el alumno ha llegado.

Otra fórmula de organización de ítems es según el objetivo (nivel cognitivo) al que se refieran. En este caso, el evaluador puede detectar el nivel de logro del alumno en cada objetivo con mayor facilidad. Dado que no suele ser compatible esta segunda fórmula de organización con la primera, el profesor, en cada caso, debe valorar qué le interesa más o qué está más de acuerdo con sus intereses al medir el aprendizaje.

d) Instrucciones

Sin duda son parte importante de una prueba. Conviene introducirlas por escrito, previamente a los elementos a que correspondan. Si hay varios grupos de ítems de distinto tipo, pueden definirse "separaciones" o "partes" de la prueba (con letras o números), cada una de las cuales llevará sus instrucciones específicas de respuesta.

Deben contener lo que pudiera definirse como "reglas del juego": tiempo disponible para responder; mecánica para registrar o anotar las respuestas y manera en que se asignarán las puntuaciones (especialmente el hecho de descontar o no, de la puntuación total, una ponderación del valor negativo de los errores).

7.2.3. Problemas

Un tipo de cuestión o reactivo que puede utilizarse en la evaluación educativa es el denominado "problema". Pocos autores caracterizan a los problemas en su calidad de reactivos o cuestiones para la evaluación de logros escolares; nosotros le dedicaremos algunas líneas que permitan aportar ciertas sugerencias para mejorar las condiciones y

consecuencias de su utilización.

De un modo general, como afirma Carreño (1977), "un problema es una proposición compleja de elementos a la que corresponden una o varias soluciones o una incógnita a despejar, atendiendo al manejo que se haga de tales elementos". De aquí que su idoneidad se halle ligada con la aplicación de principios, leyes, fórmulas, normas o criterios. Ahora bien, no sólo puede ser útil para evaluar logros, sino en otros muchos momentos del proceso de enseñanzaaprendizaje. De hecho, al programar y en ocasiones, se incluyen objetivos que implícita o explícitamente contemplan el logro como resolución de ciertos problemas; al explicar sus aplicaciones a través de la resolución de problemas; en otras ocasiones, el profesor utiliza problemas: para ejercitar el pensamiento o razonamiento del alumno, para ejemplificar procedimientos, para motivar o despertar el interés del alumno, para demostrar el uso de conceptos, etc.

Así pues, para la evaluación educativa, el problema puede utilizarse prácticamente en todos los ámbitos del aprendizaje. Veamos algunos ejemplos: resolver una ecuación matemática, calcular el coste de una aplicación publicitaria, balancear una fórmula química, determinar el valor de un parámetro en estadística, decidir qué normas legales son aplicables a un caso dado (derecho), clasificar una obra literaria dentro de un género, establecer un diagnóstico pedagógico (en educación) o un diagnóstico (en medicina), calcular la resistencia de unos materiales, etc. En todos los casos anteriores puede apreciarse un denominador común a los diferentes tipos de problemas: una secuencia ordenada de pasos o algoritmo de resolución. Es decir, para resolverlos es preciso conocer y aplicar una secuencia ordenada de pasos u operaciones interdependientes. En este sentido, con los problemas se puede medir:

-Logro de objetivos operacionales dependientes.

-Comprensión del manejo de conceptos.

-Habilidad para ejecutar operaciones implícitas.

-Obtención de una solución plausible o correcta.

Lógicamente, y bajo este prisma, la calificación o valoración de un problema ha de realizarse ayudándose de una tabla o esquema, construido por el mismo profesor, en que se especifiquen los aspectos, detalles o elementos cuya verificación es valiosa para el evaluador. Dicho de otra forma, y como afirma Carreño (1977), "cada problema debe ser fraccionado en tantos pasos como momentos demostrativos del aprendizaje puedan apreciarse, independientemente de que se llegue o no al resultado acertado" (pág. 54).

Los problemas, como vemos, pueden ser de gran utilidad para medir el aprendizaje alcanzado con relación a objetivos que reflejen un todo integrado. A menudo, sólo

presentando ese todo a conocer o a dominar, es posible evaluar objetivos de aprendizaje relativos al mismo.

7.2.4. La prueba de ensayo

Entre los exámenes de diverso tipo que se utilizan para medir el progreso del alumno en el aprendizaje, destacan las llamadas pruebas de ensayo (denominadas más comúnmente exá menes escritos), de uso generalizado en las escuelas y centros educativos. Dichas pruebas son especialmente aconsejables para medir productos de aprendizaje no reductibles a términos objetivos, esto es, productos cuyas características sean las de un sistema compuesto, de límites mayores o de índole diferente a la estricta suma de aprendizajes más simples que pudieran integrarlo (por ejemplo, capacidad para estructurar ideas, para integrar conceptos, para valorar hechos o situaciones complejas, etc.).

Siguiendo a Carreño (1977), el aprovechamiento complejo está contenido en aquellos objetivos cuyo logro consiste en producir ideas, crear formas y valorar hechos, fenómenos y procesos tales como los siguientes (columna izquierda). Sería adecuado evaluarlos a través de cuestiones o preguntas cuyas instrucciones fueran (columna derecha):

Objetivos		*Instrucciones*
a) Diseñará un experimento en el que	→	a) Diseñe un experimento ....
b) Propondrá una hipótesis que explique		b) Proponga una hipótesis que explique ....
c) Comparará la influencia que han tenido en la historia		c) Compare la influencia que han tenido ....
d) Llegará a conclusiones válidas a partir de		d) Concluya, a partir de ....
e) Analizará los efectos políticos		e) Analice los efectos políticos que acarreó .....

En las pruebas de ensayo, el alumno estructura y elabora su respuesta con una amplia libertad para: usar formas de expresión y vocabulario que crea adecuados, seleccionar y organizar sus ideas y argumentos de la manera que estime conveniente y llegar a las conclusiones y resultados que crea satisfactorios. Esta libertad suele conllevar una mayor flexibilidad en la forma y extensión de su trabajo.

Si la ventaja mayor que presentan estas pruebas es la posibilidad de estimar la consecución de objetivos complejos que no son apreciables mediante pruebas objetivas, también es verdad que, por lo general, su uso es demasiado frecuente por parte de los profesores, que al utilizarlas no ponderan la posibilidad de medir ciertos objetivos de aprendizaje a partir de otros instrumentos que no sean el examen escrito o prueba de

ensayo. Los problemas más críticos que plantean este tipo de pruebas son (Carreño, 1977):

-Tiempo y esfuerzo excesivos en la fase de revisión y calificación.

-Disparidad y falta de consistencia en las calificaciones, mediciones o puntuaciones otorgadas (tanto por diferentes examinadores, como por el mismo en momentos distintos). Subjetivismo calificador.

-Interferencia de factores que normalmente no deben participar en la evaluación, pero que en este caso alteran la apreciación del rendimiento desde distintos ángulos: presentación del material (ortografía, limpieza, claridad, etc.); pobreza o riqueza de voca bulario; previo contacto personal y relación afectiva con los examinados; habilidad para expresarse por escrito, etc.

-Falta de amplitud en la representatividad de los contenidos examinados, dejándose - en la mayoría de las ocasiones - fuera de prueba un alto porcentaje de contenidos, aspectos del programa y objetivos de igual o similar importancia que los incorporados (excesiva influencia del azar).

-Desviaciones involuntarias y frecuentes en el desarrollo del trabajo, por falta de coincidencia entre lo que el examinador y el examinado consideran que debe ser la línea del planteamiento a seguir, lo cual constituye la base de la exposición y el valor de los aspectos, conceptos y argumentos que merecen ser tratados.

-La calificación en estas pruebas no se puede circunscribir a los moldes propios de la evaluación criterial, teniéndose que atener, en cierta medida, a la mecánica de la evaluación normativa. "Téngase en cuenta que un criterio de logro exige objetividad de las conductas (y, por tanto, de las respuestas en una prueba) y en pruebas con libertad plena para elaborar respuestas - cuyo valor se asigna respecto a ciertos parámetros subjetivos del evaluador - es muy dificil decidir cuándo se han cumplido o logrado los objetivos de aprendizaje, teniendo que recurrir, al calificar, a puntuaciones intermedias, a la comparación entre varios o todos los exámenes, que equivale a una evaluación con referencia a normas".

A) Elaboración de las pruebas de ensayo

Basándonos en las ventajas de este tipo de pruebas y teniendo en cuenta las limitaciones que presentan, es posible justificar la utilización conjunta de pruebas objetivas junto a pruebas de ensayo en la medición de rendimiento. Las primeras medirían aprendizajes simples y las segundas aprendizajes complejos. No obstante, ello plantea la dificultad de provocar desconcierto en la fase de calificación, pues habrán de hacerse aditivas las calificaciones provenientes de ambos tipos de medidas - de naturaleza muy distinta -

obtenidas con relación a una misma área de aprendizaje.

De aquí que algunos autores recomienden conciliar las ventajas de ambos tipos de técnicas (pruebas objetivas y pruebas de ensayo) utilizando "preguntas de respuesta restringida". En dichas cuestiones se limita la libertad del alumno en la respuesta: fijando extensión máxima, o introduciendo criterios de logro, o perfilando los aspectos o cuestiones a responder, o circunscribiendo la forma de organizar la respuesta, lo cual ciertamente limita la creatividad. A la vez, las respuestas se aproximan lo suficiente a la objetividad de manera que se facilitan calificaciones más precisas y fiables.

Las preguntas de respuesta restringida pueden ser de extraordinaria utilidad en la mayoría de los casos en que no sea aconsejable medir a través de pruebas objetivas exclusivamente, dado que evitan la gran proporción de subjetividad que suponen las pruebas de ensayo.

A modo de orientaciones, resumimos a continuación algunas normas o consejos que pueden seguirse para construir preguntas tipo ensayo (Carreño, 1977, págs. 74-77):

-Incluir - si es factible - alguna indicación sobre el criterio de logro que el profesor vaya a emplear al calificar, a fin de que el examinado trabaje en una sola dirección (demostrativa del logro), sin perderse en inútiles parrafadas con las que cubrir distintos ángulos del área en que se desenvuelve y sin saber dónde, cómo y cuándo terminar su respuesta.

-Incluir en la pregunta las bases sobre las cuales habrá de evaluarse o calificarse la respuesta.

•Ejemplo:

Tema	"Planteamiento de los antecedentes del macartismo en la sociedad norteamericana y la posición política de sus principales líderes"
Indicaciones sobre el logro ...	Exponga las circunstancias que llevaron directamente a tal consecuencia y la ubicación ideológica común de los líderes
Bases para calificar	Se calificará la mención y articulación de dichas circunstancias antecedentes (no la descripción exhaustiva), así como la localización del aspecto ideológico que identifica o unifica a los dirigentes

-Ninguna pregunta debe tener sentidos o intenciones ocultas para el alumno (a menos que sea parte del objetivo a evaluar).

-Definir productos de aprendizaje, que se espera que el estudiante haya logrado, en términos de conducta.

-Cerciorarse de que las respuestas esperadas corresponden al logro de objetivos específicos programados.

-Asegurarse de que los aprendizajes que se tratan no son susceptibles de exploración mediante ítems de prueba objetiva (en caso contrario, deberá preferir estos últimos).

-Formule preguntas en cuya contestación se manifiesten las conductas que se buscan y no otras.

-No sorprender a los alumnos con limitaciones o condiciones de última hora referidas al tiempo para responder o a recursos que no deben ser empleados. Advertir previamente de tales límites.

B) Calificación de las pruebas de ensayo (Aproximación a la norma o al criterio)

Si las recomendaciones anteriores para la construcción de las preguntas o pruebas de ensayo han sido seguidas, se habrán conseguido márgenes aceptables de fiabilidad mediante una presentación clara, definida y delimitada de la tarea que el alumno deberá realizar.

A la hora de calificar, el profesor previamente deberá construir una tabla o esquema de corrección de las respuestas en que se destaquen los puntos importantes que haya incluido para medir el aprendizaje del alumno y la ponderación o peso que se asignará a cada parte o porción de respuesta. Asimismo reflejará en dicho esquema los puntos a sumar por el mayor o menor cumplimiento en cada examen de aquellos criterios o características que "pesan" en las respuestas: originalidad, estructuración del tema, adecuación de ejemplos, dominio de lenguaje apropiado, coherencia en el manejo y encadenamiento de argumentos, etc. También incluirá en dicho esquema puntuaciones que habrán de restarse en caso de presentarse información no pertinente, redundante, falsa... (siempre que se haya advertido de tal sustracción al alumno previamente).

La revisión de exámenes "habrá de realizarse por partes (respuestas, puntos tratados, aspectos del trabajo, etc.) calificando cada parte en todos los exámenes antes de pasar al siguiente... Con este mecanismo se corrige en gran medida la influencia que unas respuestas pueden tener sobre otras del mismo alumno, y se tiene más reciente y presente el criterio con el cual se está calificando" (Carreño, 1977, pág. 79).

De entre los procedimientos a emplear para el proceso mismo de la calificación pueden destacarse tres modelos que siguiendo a Carreño (1977) son: el primero - de aproximación a la norma - coincide con los principios de la didáctica tradicional; el

segundo - de aproximación al criterio - se acerca más a lo establecido por la tecnología educativa moderna; siendo el tercero - peso relativo - una simplificación del proceso evaluativo a la que cabe recurrir cuando se trate de exámenes cuyos resultados no tengan gran trascendencia en la vida académica del alumno. Sucintamente y siguiendo a dicho autor podrían reflejarse así (págs. 79-82):

a) Aproximación a la norma

1. Se elabora un modelo óptimo con la o las respuesta/s más completas, asignándose valor en puntos a cada porción o aspecto según su grado de dificultad y la importancia del aprendizaje que demuestran. Se asignan máximos a cada parte.

2. Se revisan los exámenes o pruebas, asignando puntos a cada porción o aspecto según se aproxime a las características del modelo con el máximo o tope definido en el punto anterior.

3. Se suman o restan puntos por factores independientes que deban ser evaluados (bases para calificar).

4. Se totaliza la puntuación obtenida y se interpreta normativamente (baremos).

b) Aproximación al criterio

a) Se elabora el modelo tomando como base el mínimo necesario de elementos que deben incorporarse y las formas más elementales aceptables, para considerar logrado cada objetivo programado en la prueba.

b) Se clasifican los aspectos o partes de las respuestas según pertenezcan al logro de uno u otro de los objetivos evaluados.

c) Se revisan los exámenes o pruebas, asignando en cada aspecto si se cumplen o no los requisitos establecidos en el modelo.

d) Se agrupan los logros parciales y se determina si son suficientes o no para considerar globalmente logrado cada objetivo.

e) Se otorgan las notas en función del número y jerarquía de los objetivos logrados con el total de objetivos medidos.

c) Peso relativo

1. Se elabora, más que un modelo, una relación sucinta de los puntos o aspectos que serán evaluados.

2. Se define una base de clasificación con varias categorías (3, 4 o 5) que representen distintos grados de calidad que a grandes rasgos puedan identificarse con:

   •TRES: Malo/regular/bueno.

   •CUATRO: Deficiente/suficiente/bueno/muy bueno.

   •CINCO: Muy deficiente/insuficiente/medio o suficiente/bueno/muy bueno.

   Estos adjetivos han de tener pleno valor discriminatorio para el evaluador. Si los anteriores adjetivos no le ayudan, debe "inventarlos" él mismo.

3. Se preparan espacios físicos (huecos) donde ir acumulando las pruebas según sean revisadas (se facilita con un rótulo que indique la categoría valorativa).

4. Se revisan las pruebas o exámenes, clasificándolos en una de las categorías definidas (paso 2) y se acumulan en uno de los espacios creados en el paso anterior. Se opta por corregir pregunta a pregunta, estos pasos se repetirán para cada pregunta.

5. En caso de varias clasificaciones, con las notas acumuladas se llega a una clasificación definitiva ("promediando" dichas notas por "mayoría" o predominio de alguna sobre las demás).

6. Se comparan las notas clasificadas, con la escala de clasificación que vaya a utilizarse y se establecen las equivalencias del caso conforme al significado de cada categoría y cada grado de la escala asignándose las calificaciones que correspondan (por ejemplo, de cuatro categorías, tomado de Carreño, 1977, en el cuadro 7.12).

CUADRO 7.12. Ejemplo de correspondencias entre categorías de clasificación, significado de cada categoría y valores de una escala numérica de calificación

Categorías	*Significado*	*Escala*
Deficiente	Se considera insuficiente para aprobar	0, 1, 2, 3, 4
Mediano	Cumple apenas los requisitos mínimos exigibles	4, 5, 6
Bueno	Demuestra con suficiencia y claridad el logro de los objetivos	6, 7, 8
Muy bueno	Demuestra ampliamente el logro de los aprendizajes, haciendo gala de coherencia, organización y profundidad	8, 9, 10

8

La práctica evaluadora en el centro escolar

La evaluación educativa - como hemos defendido en los capítulos iniciales - es una actividad o proceso de identificación, recogida y tratamiento de datos sobre elementos y hechos educativos con el objeto de valorarlos, en primer término y, sobre dicha valoración, tomar decisiones.

Una tarea como la evaluadora exige una cuidadosa planificación que vaya desde la especificación de los juicios que deben emitirse y las decisiones que deberán tomarse hasta la formulación de dichos juicios y decisiones, que deberán resumirse y darse a conocer a los interesados: profesores, padres, alumnos y, en general, comunidad educativa. Entre el principio y el final del proceso deberán cubrirse una serie de etapas, tales como: recogida de información previa, definición de objetivos educativos, determinación de los sistemas de referencia a utilizar para cada objetivo, construcción y selección del procedimiento o técnicas para registrar o medir variables relevantes, aplicación de pruebas, etc.

La planificación de la evaluación debe ser, por lo tanto, minuciosa, y debe implicar al conjunto de profesores de un centro escolar o, cuando menos, al conjunto de profesores de un ciclo, o de un curso específico.

Una posibilidad defendida por diversos autores es la elaboración de un programa general de evaluación para todo el centro escolar sobre la base de pruebas tipificadas aplicadas a todos los sujetos. Ello presenta ventajas e inconvenientes. Antes de exponer el citado proceso tipo de evaluación, vamos a analizar esta posibilidad brevemente.

8.i. Planificación general de la evaluación en un centro escolar

Un programa general de evaluación consiste, básicamente, en un procedimiento unitario de evaluación del proceso de los alumnos de un centro escolar en las áreas cognitiva y no cognitiva. Suele realizarse a través de pruebas estandarizadas. La decisión de realizar dicho programa general deberá estudiarse en el contexto de cada centro; en unos podría ser posible y resultar beneficioso, en otros quizá no lo fuera o no resultaría adecuado.

Es obvio que la finalidad de un programa general es presentar a los docentes información que les ayude a preparar un ambiente de aprendizaje más idóneo para los

alumnos. Si logra este objetivo será satisfactorio, si por el contrario no logra mejoras en el ambiente de aprendizaje, habrá de considerarlo inadecuado, no ajustado a las necesidades de los alumnos.

Las ventajas de un programa general - siguiendo a Beggs y Lewis, 1979 - pueden ser varias:

-Ayudar al profesor a evaluar la calidad del currículo de materias escolares.

-La observación de los resultados de la aplicación simultánea de pruebas a todos los alumnos puede permitir a los profesores determinar si los alumnos de diferentes etapas poseen capacidades y conocimientos que deberían poseer en ese momento o etapa escolar, así como si se van alcanzando los objetivos del currículo o si éste precisa cambios.

-El profesor puede también encontrar puntos fuertes y débiles y prever, en consecuencia, una acción individualizada (fines diagnósticos).

-Facilitar la agrupación de los alumnos para actividades instructivas específicas.

-Determinar la satisfacción de los alumnos dentro del centro, utilizando tests o pruebas de actitudes.

-Informar a la comunidad educativa sobre el rendimiento de los alumnos en la actividad escolar. En este caso, el centro debe asegurarse de que la comunidad educativa interpretará correctamente los datos y otros aspectos críticos incluidos en el informe. Deberá evitarse utilizar estos datos como única fuente para evaluar al profesor, a los alumnos o al propio centro en su actividad docente.

Ciertamente, este programa general de evaluación no excluye las actuaciones evaluadoras propias de cada profesor en su grupo concreto de alumnos, ni puede sustituir a la evaluación progresiva, base fundamental de una evaluación formativa y no meramente sumativa.

Un programa general como éste exige como primera medida un amplio consenso entre los profesores de un centro. Posteriormente se debería formar un comité de profesores y directivos del centro que definirá: las líneas maestras, momentos de las aplicaciones, la preparación del personal aplicador e interpretador de los resultados y las pruebas o tests de medida (internos o externos) que se utilizarán.

Obviamente, todo lo anterior lleva tiempo, de aquí que se aconseje disponer de una persona cuya principal responsabilidad sea la supervisión efectiva del programa. Dicha persona, que deberá ser un especialista en técnicas de medición y recogida de datos (un pedagogo, por ejemplo), no tomará decisiones fundamentales como las anteriores (tests

que se aplicarán, momentos, etc.), pues esta tarea debe quedar en manos de los profesores que forman parte del comité; debe sugerir y recomendar, al objeto de que se facilite la toma de decisiones en el comité. Una vez tomadas dichas decisiones, debe hacerlas efectivas supervisando la aplicación del programa hasta sus últimas consecuencias.

Elegido el comité de profesores y seleccionado su director o coordinador, los pasos principales serían (Beggs y Lewis, 1979):

-Establecimiento de los objetivos del programa. Definición de objetivos educativos, momentos de recogida de datos, definición de la importancia relativa de los contenidos, determinación de los instrumentos o pruebas a utilizar (externos o internos, de referencia criterial o normativa), determinación de las funciones de la evaluación (de control, diagnóstico, etc.).

-Establecimiento de los procedimientos de aplicación. A través de sesiones de planificación con todos los profesores implicados, se decidirán: fechas, duración de las pruebas, información a dar a los alumnos, etc.

-Establecimiento del procedimiento para elaborar informes de resultados a profesores, padres, alumnos y comunidad educativa en general. Es especialmente importante informar a cada alumno de su desarrollo particular a través de perfiles acumulativos que informen de su progreso. Sin duda, esta información es la más relevante para el alumno, al permitirle compararse consigo mismo.

-Establecimiento de procedimientos para exámenes auxiliares que podrían ser necesarios para algunos alumnos o grupos, con posterioridad a la fase general de aplicaciones. Este tipo de pruebas será especialmente importante cuando quieran utilizarse los resultados de estas pruebas para otorgar calificaciones y sus resultados no coincidan con los datos observados personalmente por cada profesor respecto de sus alumnos.

Si la finalidad de un programa como el descrito es definir lo que conocen y saben los alumnos y su motivación, así como una información realista sobre los que no poseen esos aspectos, el programa no termina con la aplicación de las pruebas y la redacción de informes; deberán interpretarse los resultados por parte de cada profesor para intentar modificar el ambiente escolar de los alumnos en la línea de ayudarles a una mejor comprensión de los objetivos buscados por el programa general. Como afirman Beggs y Lewis (1979), "si el profesor no permite la posibilidad de una evaluación crítica de los resultados y su informe individual dentro del ambiente escolar, hay que poner en duda la utilidad y fundamentos mismos del programa general. No se debería establecer un programa general de exámenes, si los tests no se utilizan para mejorar el ambiente de aprendizaje de todos los que se encuentran dentro de la situación escolar" (pág. 240).

8.2. Pasos de un proceso tipo de evaluación

A modo de síntesis ofrecemos una secuencia de pasos de un proceso tipo de evaluación en un centro escolar. Dichos pasos integran, de alguna manera, los aspectos sustantivos definidos en los capítulos anteriores.

Para definir esta secuencia de pasos, seguimos básicamente a Tenbrick (1981). Dicho autor estructura en diez pasos el proceso de evaluación. Para no ofrecer una secuencia en un número distinto de pasos, nosotros utilizaremos también diez pasos, que no coinciden necesariamente con los señalados por Tenbrick, aunque siguen sus orientaciones. Las principales fases del proceso de evaluación podrían resumirse en cuatro: planificación, selección y construcción de instrumentos, recogida de datos y evaluación en sentido estricto. La figura 8.1 explicita estas cuatro fases en los diez pasos ya señalados.

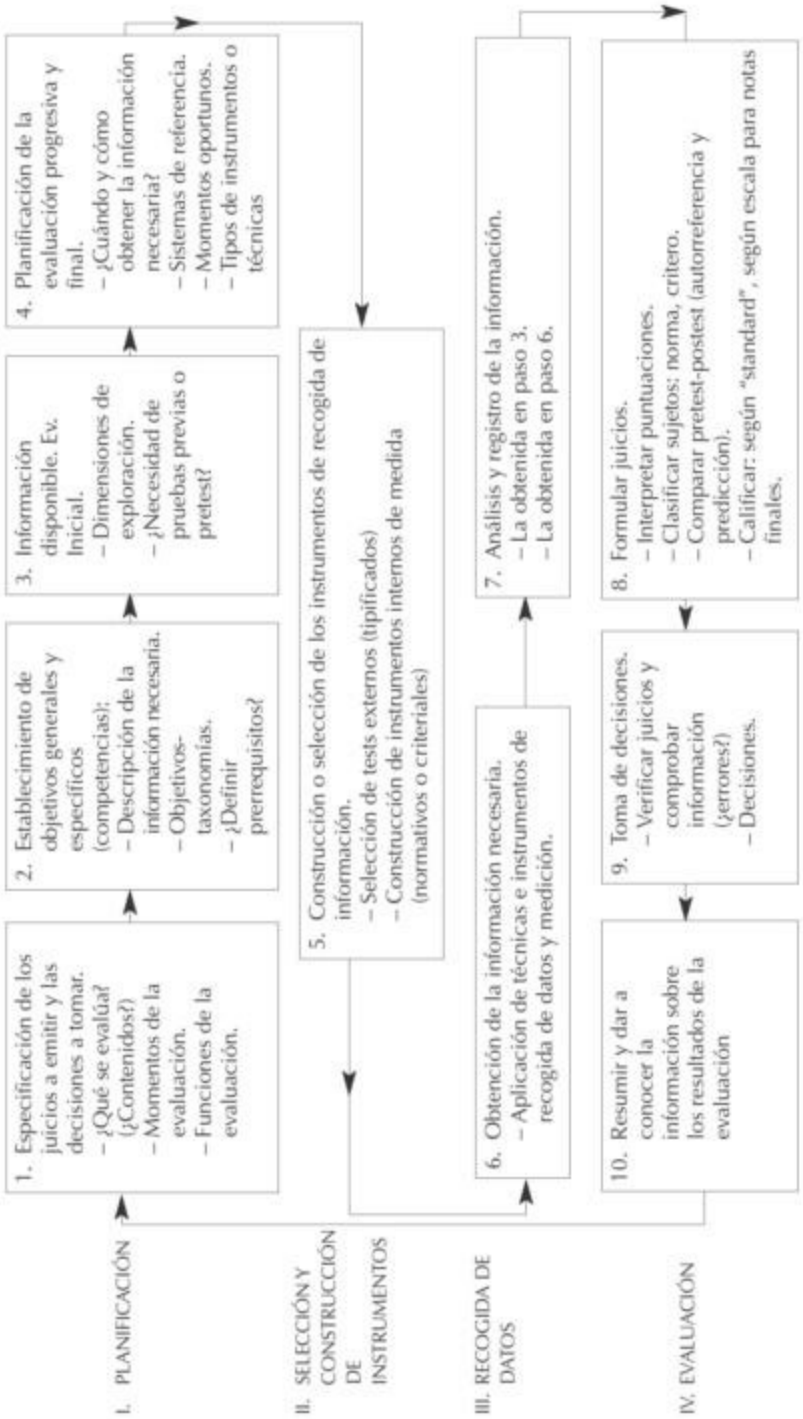


Figura 8.1. Proceso tipo de evaluación. Fases y pasos.

8.2.1. Fase de planificación

La importancia de esta fase no es preciso exagerarla. Veamos sus pasos.

1. Especificación de los juicios a emitir y las decisiones a tomar

Implica reconocer, identificar y especificar las áreas de decisión importantes. Sin dicha

especificación es dificil precisar qué clase de información necesitaremos. Como señala Tenbrick (1981), "a menudo, las decisiones que se deben tomar determinan antes y después los juicios necesarios para tomarlas. Con todo, el profesor, al pensar en una tarea evaluativa dada, algunas veces se centrará primero en los juicios que debe formular, en la clase de decisiones para la que pueden servir los juicios" (pág. 23). En todo caso, la relación entre juicio y decisión es estrecha, de manera que es dificil pensar en lo uno sin lo otro.

Los juicios y decisiones pueden ser de muy diversa índole, dependiendo de: los contenidos a evaluar, la función que se asigne a la evaluación (formativa o sumativa; diagnóstica, predictiva, de control, etc.), el objetivo de la decisión (administrativo, orientador, de enseñanza o de investigación), importancia de las decisiones (reversibilidad de las mismas, número de individuos afectados, tiempo que durará el efecto de la decisión, importancia que se asigne a los resultados de aprendizaje).

También es necesario preguntarse cómo se debe formular un juicio. La respuesta la hallaremos si prevemos el sistema de referencia a utilizar (normativo, criterial o de autorreferencia). Los juicios, pues, podrán hacer referencia a la norma, a un criterio (nivel de dominio en un continuo establecido) o a sí mismos (juicios sobre progresos y cambios).

Tenbrick (1981) utiliza tres variables para distinguir un tipo de juicio de otro:

a)Factor tiempo (estimaciones presentes o predicciones futuras).

b)Factor sistema de referente utilizado (autorreferencia, norma y criterio).

c)Factor tipo de comparación (de dominio, si el referente es una puntuación límite o de desviación, si se determina la desviación del referente a lo largo de una escala o contínuum de realización).

Combinando las categorías o niveles de tres variables pueden darse 12 tipos distintos de juicios (2 x 3 x 2 = 12). La figura 8.2, tomada de dicho autor, refleja estos doce juicios posibles.

2. Establecimiento de objetivos generales y específicos

Una vez especificados los juicios y decisiones que deberán darse al final del proceso evaluativo, estamos en condiciones de describir la información que será necesaria para ello. Cuanto más claramente describamos dicha información, más fácilmente podremos preparar todo lo necesario para obtenerla.

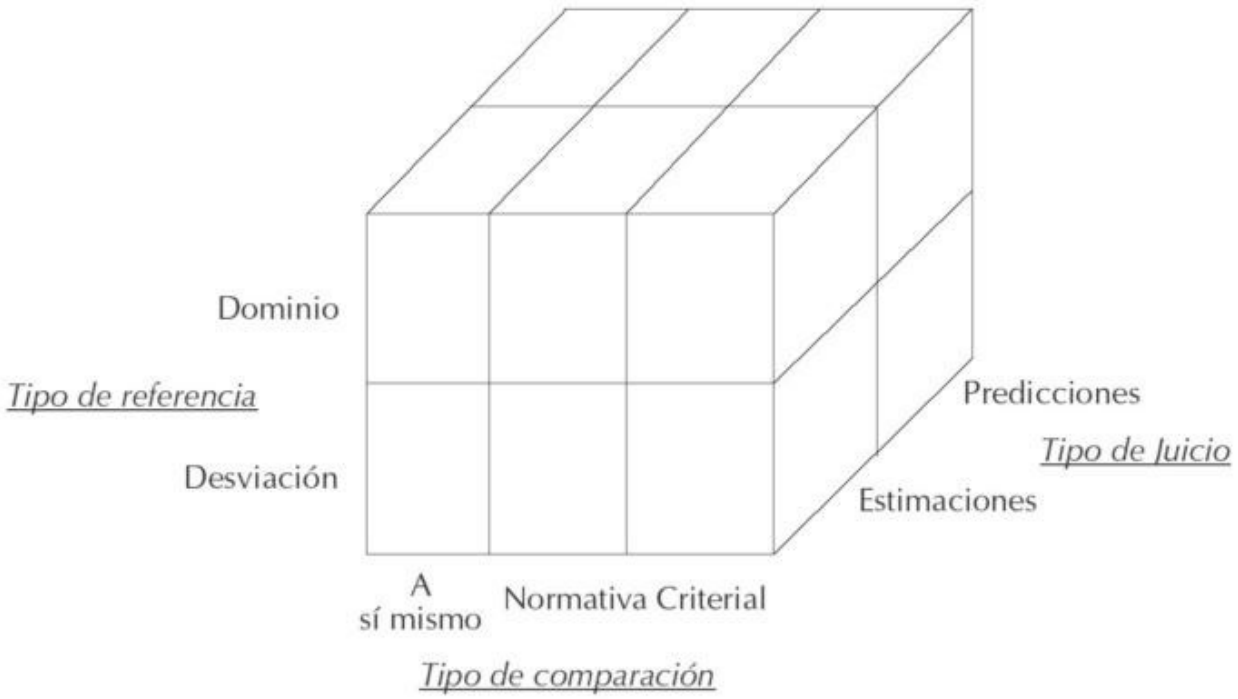


Figura 8.2. Doce juicios posibles, resultantes de la combinación de los factores (tiempo, sistema de referencia y tipo de comparación).

En este paso del proceso, como señala Tenbrick (1981), será preciso responder a cuatro preguntas:

a)¿Sobre qué es la información?

b)¿Qué clase de comportamiento se va a observar?

c)¿Hasta qué punto debe ser especificada la información?

d)¿Hasta qué punto debe ser exacta la información?

La primera cuestión deberá responderse en términos de contenidos de aprendizaje y tipo de respuestas que se exige al alumno. La segunda cuestión podrá responderse en términos de objetivos educacionales. Las taxonomías de objetivos pueden ayudarnos a especificar el tipo de comportamiento que será preciso observar. La respuesta a la tercera cuestión estará claramente relacionada con los juicios y decisiones para los que se utilizará la información (número y tipo de juicios). Deberá responderse en función de lo especificado en el primer paso del proceso. A la última cuestión se responderá en términos de precisión necesaria en la recogida de información. Si los juicios y decisiones a tomar son o pueden ser "graves" (irreversibles, gran número de sujetos afectados, etc.) será necesario exigir mayor precisión y exactitud en la información a recoger. Ello

implicará estimar el coste de la recogida de datos y medida de variables, así como planificar cuidadosamente las técnicas e instrumentos a utilizar para recoger la información.

Resumiendo, en este paso del proceso evaluativo deberán definirse:

-Objetivos generales y específicos (ayudarse de taxonomías).

-Las tablas de especificaciones que relacionen objetivos y contenidos de aprendizaje.

-Contínuum de realización del alumno.

-Cada comportamiento en términos observables (objetivos operativos).

-Orden significativo de los comportamientos (orden secuencial, orden jerárquico o taxonómico, o bien orden de desarrollo).

Finalmente, será preciso definir los requisitos de entrada o prerrequisitos (si los hay) que el alumno deberá dominar o conocer para incorporarse con posibilidades al proceso de enseñanza-aprendizaje.

3. Información disponible. Evaluación inicial

Establecidos los objetivos educativos y los prerrequisitos correspondientes al proceso educativo diseñado, en este momento intentaremos recoger la información que pudiera estar ya disponible sin necesidad de aplicar técnicas e instrumentos de recogida de datos o definir los instrumentos o técnicas de recogida de información que será preciso aplicar para realizar una evaluación inicial, tanto de prerrequisitos, como de variables relevantes en el proceso educativo a evaluar.

Ello implica:

a)La definición de las dimensiones básicas de exploración inicial (datos personales, ambientales, familiares, datos médicos, datos psicológicos y antecedentes académicos).

b)Recogida de información ya disponible sobre dichas variables o dimensiones. Dicha información puede encontrarse en:

1.Notas personales de profesores anteriores.

2.Archivos del centro (expedientes).

3.Archivos del gabinete ¿e orientación.

4.Notas del personal auxiliar de la escuela.

5.Notas de los padres.

En dichas fuentes de información habrá que buscar datos sobre las variables relevantes que hemos definido. Habrá que cuidar la clasificación de dicha información, pues no todos los datos recogidos tendrán igual fiabilidad, relevancia y utilidad. Puede también ser posible que haya datos contradictorios. Por todo ello, el profesor debe valorar la información que recoge, diferenciando: informaciones, medidas, apreciaciones, hechos, etc., así como el método o procedimientos que se utilizaron para recoger la información (observación directa o indirecta, entrevistas, cuestionarios, tests, escalas, etc.).

c)Selección de técnicas o instrumentos para recoger información sobre variables o dimensiones relevantes de evaluación inicial de las que no se tienen datos, o bien, dichos datos que posee no son fiables o útiles.

En ocasiones, nuestra búsqueda de información puede haber sido infructuosa o insuficiente en sus resultados. Puede ser preciso recoger alguna otra información relevante sobre prerrequisitos o sobre variables relevantes de cara a la evaluación progresiva y final. Será necesario aplicar instrumentos disponibles o diseñar técnicas o procedimientos que nos permitan recoger esa información.

En todo caso, el profesor o equipo de profesores, en cada caso, deberá valorar la información de que dispone y la necesidad de recoger más información de cara a la evaluación inicial, dependiendo de: el momento de desarrollo del sujeto, la importancia de la información que se desconoce, la "gravedad" de los juicios y decisiones a tomar en el proceso evaluativo y los objetivos mismos de la evaluación. Si se pretende, por ejemplo, determinar el progreso individual de un sujeto tras un programa educativo, es evidente que la información sobre los rendimientos previos o dominio de destrezas básicas puede ser muy relevante, siendo necesario, en tal caso, medir a través de pruebas previas - pretests - dichas variables.

4. Planificación de las evaluaciones inicial, progresiva y final

Es un hecho frecuente, que en la mayoría de las ocasiones, la información que se necesita no estará disponible, de aquí la importancia de planificar cuidadosamente el momento y la forma de recoger información, asegurando su fiabilidad y validez.

En este paso, debemos decidir el cómo (la forma) y el cuándo (los momentos) recoger información.

La respuesta a la primera cuestión implica tomar decisiones de selección o

construcción de técnicas e instrumentos de recogida de información. Es obvio que decidir qué técnica o qué instrumento utilizar para cada información necesaria dependerá de qué información se trate. El segundo bloque de este libro le ayudará a decidir qué instrumentos o técnicas seleccionar (capítulos 6 y 7). En la página 91 se encuentra una clasificación de las principales técnicas e instrumentos de recogida de datos y medida, que resultará al lector de gran utilidad. Dicha clasificación diferencia entre:

a)Técnicas de recogida de datos. Observación (directa o indirecta) y encuesta (cuestionarios, entrevistas y técnicas sociométricas).

b)Instrumentos de medición. De aptitudes (tipificados), de personalidad, intereses, actitudes, adaptación, etc. (también tipificados) y de rendimiento y pedagógicos (externos, tipificados e internos, construidos por el profesor o centro).

Una vez que se ha decidido qué información se necesita (paso 2), de qué información se dispone (paso 3) y qué instrumentos se utilizarán para obtenerla, es preciso decidir cuándo obtenerla. De nuevo, la respuesta a esta cuestión se debe tomar sobre:

a)El tipo de información necesaria.

b)Objetivos y funciones de la evaluación.

c)Los tipos de juicios y decisiones que se van a formar.

Un ejemplo tomado de Tenbrick (1981) puede ayudarnos. "Si se necesita información sobre el rendimiento máximo de un estudiante, éste debe haber tenido una oportunidad de alcanzar su `máximun' y debe saberlo por adelantado para poderse preparar a hacerlo lo mejor que pueda. Si se necesita información sobre rendimiento típico, se debe obtener sin previo aviso. Si estamos tratando de formular juicios formativos, necesitamos conseguir información durante el aprendizaje, no después de que se ha realizado. Si queremos una medida de logro de objetivos últimos, necesitamos esperar hasta que cada alumno haya tenido una oportunidad de alcanzar esos objetivos" (pág. 146).

8.2.2. Fase de construcción y selección de los instrumentos de recogida de información

Lo que sí es importante (paso 5) es anticipar cuándo comenzará y terminará la instrucción, al objeto de tener tiempo suficiente para seleccionar o construir los instrumentos que se necesitarán para la recogida de información. Tanto si se trata de exploración inicial, como de evaluación progresiva y final, los instrumentos tipificados tardan en llegar al centro desde la casa editora de los mismos; los instrumentos internos y de registro de datos, elaborados por el profesor o equipo de profesores, requieren, normalmente, mucho tiempo de construcción si ésta se realiza adecuadamente.

De hecho, si se trata de recoger información a través de instrumentos de registro de

datos, será preciso decidir cuáles se van a utilizar (observación directa o indirecta, instrumentos de encuesta) para obtener información sobre qué variables, conductas o comportamientos especificados en los pasos 2 y 3.

Si se trata de tests y pruebas externas (tipificadas), habrá que seleccionar de entre los disponibles.

Si se trata de construir tests internos para medir variables de rendimiento, habrá que diseñarlas de acuerdo a:

-Los objetivos de la educación especificados.

-El tipo de referencia a utilizar para medir cada objetivo.

-El tipo de instrumento elegido (pruebas objetivas, problemas, pruebas de ensayo, etc.).

8.2.3. Fase de recogida de datos

En esta fase se trata de aplicar las técnicas o instrumentos de recogida de datos y medición (paso 6), seleccionados o diseñados y de analizar y registrar la información obtenida (paso 7) de manera que estemos en disposición de formular juicios y tomar decisiones evaluativas (pasos 8 y 9).

Paso 6. Aplicación de técnicas e instrumentos de recogida de información y medida

Tenbrick (1981) afirma que el obtener información práctica depende en gran parte de la oportunidad. Por ello, antes de intentar obtener información evaluativa, un profesor debe intentar hacerse dos preguntas: ¿estoy preparado?, ¿están preparados los alumnos? (pág. 156). La respuesta a estas cuestiones no es baladí. Sólo si respondemos afirmativamente estaremos en condiciones de asegurar la validez aparente de nuestros procedimientos.

Si se trata de utilizar instrumentos de observación o encuesta, habrá que asegurarse de conocer bien dichas técnicas y de prestar gran atención a su aplicación. Para utilizar dichas técnicas, como medio valioso de recogida de información, deben seleccionarse los mejores instrumentos disponibles y estar familiarizado con su empleo, observar constantemente, verificando la información obtenida con otras de que se disponga y seguir las instrucciones que para cada técnica se sugieren.

Si se trata de aplicar tests (externos o internos), habrán de seguirse cuidadosamente las instrucciones que contenga el manual de la prueba (tests externos) o bien elaborar instrucciones que garanticen las mismas condiciones para todos los sujetos (tests internos o construidos por el centro).

También deben puntuarse dichas pruebas, esto es, determinar para cada sujeto las puntuaciones o calificaciones resultantes de las aplicaciones.

Paso 7. Análisis y registro de la información necesaria

Después de obtener la información (paso 6) o haber recogido la información relevante que se hallaba disponible (paso 3), debe procederse a analizarla (contextualizarla o transformarla en información útil) y registrarla (en fichas, tablas o medios que faciliten su valoración).

Como venimos afirmando en capítulos anteriores, la información obtenida ha de ser útil y lo será si sabemos en qué grado es fiable y válida por un lado y por otro, si se halla contextualizada, esto es, si se sabe qué significa, de quiénes procede y cómo se obtuvo. La mayoría de los autores coinciden en esta afirmación: la información, por sí sola, es inútil. Las puntuaciones brutas, crudas, las observaciones, anécdotas, notas de entrevista, etc., no tienen sentido si no se contextualizan (Pophan, 1980; Tenbrick, 1981; etc.). Este último autor afirma que quien pueda contestar a las siguientes preguntas ha obtenido una información que le será útil al formular juicios y tomar decisiones (pág. 172):

1. ¿Sobre qué o quién es la información? (sujeto).

2. ¿Quién obtuvo la información? (observador o aplicador).

3. ¿Qué instrumentos (si es el caso) se usaron para obtener la información?

4. ¿Cuándo se obtuvo la información? (día, fecha, hora).

5. ¿En qué condiciones se obtuvo la información?

6. ¿Es esta información aplicable a alguien más?

7. ¿Está la información relacionada con cualquier objetivo de instrucción o logros de aprendizaje?

8. Si la información es una puntuación, ¿qué tipo de puntuación es bruta, cruda o elaborada, transformada?

9. ¿Hay alguna evidencia de su fiabilidad o de su validez?

Una vez contextualizada la información, es preciso registrarla en algún tipo de medio físico (ficha, tabla, cuaderno, etc.) que permita acumularla, de forma que se puede hacer una consulta fácil, clara y comprensible de cara a la formulación de juicios y a la toma de decisiones evaluativas.

Normalmente, cada información recogida requiere un tipo distinto de medio para resumir y contextualizar la información que ofrece:

a)Tests o pruebas tipificadas (estandarizadas). Los protocolos u hojas de respuestas normalmente tienen un espacio reservado para obtener la puntuación bruta total, otro espacio para cada puntuación parcial y otro espacio para situar la puntuación transformada o tipificada (centiles, puntuaciones típicas, etc.). Asimismo, ofrecen también, cuando es relevante, un gráfico o perfil que permite visualizar la ejecución personal de cada sujeto.

b)Tests o pruebas internas. El profesor o centro debe construir una ficha que permita sintetizar la información que cada sujeto ha obtenido en cada parte de la prueba o en el total de la misma y también cuantos datos relevantes sea necesario incluir para contextualizar dicha información (fecha, curso, datos personales, materia, tipo de prueba, máximos y mínimos, datos técnicos de la prueba, tales como: fiabilidad, baremos, validez, etc., si ha lugar a ellos). Elaborada dicha ficha bastará rellenar una para cada alumno.

c)Datos de observación.

-Los datos de escalas de estimación y listas de control son fácilmente resumibles en puntuaciones por conductas o comportamientos observados.

-También los datos anecdóticos son fácilmente sintetizables en las propias fichas o en el anecdotario.

-Las calificaciones en trabajos y escalas de producción pueden ser registradas, junto a las anteriores, en el expediente acumulativo de cada alumno. Normalmente, la información sobre cada trabajo particular de un alumno puede resumirse en muy pocas formulaciones o valoraciones.

d)Datos de instrumentos de encuesta. Los datos procedentes de cuestionarios y entrevistas pueden también resumirse en unas cuantas formulaciones archivables, junto a los datos anteriores, en el expediente acumulativo del alumno.

Los datos sociométricos son más difíciles de sintetizar, dado que se precisan más referencias para situar a un alumno en un grupo de clase y por la mayor dificultad interpretativa de las propias técnicas sociométricas. En todo caso, datos referidos al grupo al que pertenece un alumno particular pueden sintetizarse en sociogramas y otros gráficos y tablas que reflejen frecuencias y porcentajes de aceptaciones y rechazos.

En resumen, es aconsejable abrir una carpeta o registro acumulativo de datos para

cada alumno particular. En dicha carpeta se irán acumulando los resultados de pruebas, tests, escalas, cuestionarios, calificaciones de exámenes o trabajos, etc. A la hora de una evaluación final, dicha información puede sintetizarse en tablas o gráficos individuales o de grupo, que permitan visualizar más fácilmente el rendimiento, logro y progreso de cada alumno con relación a sí mismo, al grupo al que pertenece o bien con relación a diversos dominios de aprendizaje definidos. Medios como los anteriores y los siguientes pueden facilitar la síntesis de información evaluativa, los juicios y decisiones evaluativas y el resumen de información posterior facilitable a los interesados (alumnos, profesores, padres y comunidad educativa en general):

-Registro acumulativo de evaluación (RAE).

-Extracto del registro personal de alumno (ERPA).

-Acta de evaluación final.

-Libro de escolaridad.

-Boletín informativo a la familia.

8.2.4. Evaluación (formulación de juicios, toma de decisiones e informes)

La última fase del proceso implica la formulación de los juicios, la toma de decisiones y la emisión de informes a los interesados en conocer y valorar los resultados del proceso educativo.

a)Evaluar, como venimos diciendo, es juzgar (paso 8). Todos los pasos enunciados hasta aquí tienen como objetivo ayudar y facilitar juicios evaluativos. Las decisiones pedagógicas que toman los profesores y responsables de la educación se apoyarán en estos juicios, lo cual convierte a la tarea de formular juicios en una fase de excepcional importancia.

Los juicios que deben formularse y que se hallarán especificados en el primer paso del proceso pueden ser muy distintos: diagnóstico de problemas, calificación de exámenes, predicción del éxito futuro, asignación de notas parciales o finales de curso, etc.

Como señala Tenbrick (1981), para formular juicios conviene recordar los tres factores que determinan cómo se forma un juicio: a) ¿es una estimación o una predicción?; b) ¿qué tipo de referente se usa?; c) ¿es un juicio de dominio o de desviación? Cada uno de estos factores influye en el modo en que se hace un juicio y, hasta cierto punto, determinan para qué tipo de decisión se puede utilizar (pág. 187).

A modo de síntesis, los principales juicios que pueden formularse son:

1. Interpretación de puntuaciones de tests externos o tipificados.

2. Hacer estimaciones. Dichas estimaciones pueden hacerse por referencia a la norma, o por referencia a un criterio.

3. Hacer predicciones de éxito futuro en materias, contenidos o habilidades relacionadas con las variables objeto de medida y evaluación. La base estadística de estas predicciones son las ecuaciones de regresión simple o múltiple (dependiendo de si se predice una variable - criterio - sobre la base de una o más variables predictoras).

4. Calificación y clasificación de sujetos sobre la base de puntuaciones obtenidas en tests internos. Tenbrick (1981) diferencia, dentro de este tipo de juicios, varias posibilidades:

-Clasificar en una curva (juicios de desviación con respecto ala norma).

-Notas, como porcentaje de puntuaciones correctas (juicios de desviación respecto a un criterio).

-Notas, como percentiles (juicios por referencia a la norma).

-Calificar en un "contínuum" de realización (juicios por referencia a un criterio).

-Calificar según un "standard" (juicios de dominio).

-Calificar en una escala "standard" (juicios de desviación, respecto de un dominio).

-Calificar para libros de notas (juicios últimos).

-Juicios en base a su comportamiento típico.

-Juicios por referencia a sí mismo (autorreferidos).

b) Evaluar es tomar decisiones, sobre la base de los juicios establecidos (paso 9). Como suele afirmarse, no hay razón más importante para evaluar que ésta: tomar decisiones, tomar las mejores decisiones posibles o como dicen algunos autores "elegir entre decisiones alternativas" aquella o aquellas más satisfactorias. Siguiendo a Stufflebean y otros (1971) y a Tenbrick (1981), y de forma resumida, los pasos fundamentales a seguir en el proceso de toma de decisiones son:

1. Verificar juicios.

1.1. Comprobar la información sobre la que se basa el juicio.

1.2. Obtener evidencia empírica de la verdad del juicio.

2.Tomar decisiones.

2.1.Considerar las causas más frecuentes de decisiones incorrectas (formulación de juicios inadecuados, las alternativas consideradas son muy pocas, errores al considerar las consecuencias de las decisiones, no tener en cuenta el sistema de valores, empleo de juicios defectuosos).

2.2.Pasos de la toma de decisiones.

2.2.1.Especificar el objetivo (determinar qué se quiere que consiga una acción).

2.2.2.Identificar las posibles alternativas (y sus resultados más probables).

2.2.3.Considerar las consecuencias de cada tipo de acción y sus probables resultados.

2.2.4.Escoger la mejor alternativa.

c)Emisión de los informes de evaluación. Este sería el último paso del proceso (paso 10). En él se sintetizarán los resultados del proceso evaluativo para darlos a conocer a los múltiples interesados en ellos: estudiantes, padres, directores, orientadores y comunidad educativa en general.

Es evidente que un paso previo a la elaboración de dichos informes será la síntesis - por parte del profesor - de los resultados de la evaluación. Para ello, el profesor contará con una carpeta individual de cada alumno en la que figuren los datos que ha ido recogiendo a lo largo del proceso, así como los juicios y decisiones que han sido elaborados en los pasos anteriores.

Será también de gran utilidad para el profesor conocer con la mayor precisión posible:

-Las disposiciones legales que afectan a la evaluación (tanto de modo general para el Estado Español, como para cada Comunidad Autónoma) correspondientes al nivel de enseñanza en que realiza sus tareas docentes.

-Los documentos de evaluación específicos para el nivel de enseñanza de que se trate (ERPA, RAE, acta de evaluación, libro de escolaridad, boletín de información a la familia, etc.).

Durante décadas, los profesores han informado del progreso de los alumnos a los interesados, utilizando para ello puntuaciones que indicaban el nivel de rendimiento. No obstante, el boletín de notas y el sistema de calificación tradicional han sido cuestionados

en numerosas ocasiones a lo largo de la historia de la educación. El principal argumento utilizado ha sido que los libros o boletines de notas no informan eficientemente del progreso del alumno. No obstante, los libros de notas y las puntuaciones siguen utilizándose: ¿por qué? Tenbrick (1981) señala dos razones por las que hoy se siguen usando boletines de notas y puntuaciones. En primer lugar, porque hay una necesidad real de una indicación precisa, fácilmente archivable, sobre logros y progresos del alumno (sencillez que ofrecen las puntuaciones); a la par, estas descripciones numéricas del progreso han sido consideradas como predictores particularmente eficientes de realizaciones escolares futuras. Una segunda explicación de este uso continuado de las calificaciones puede ser que los expertos en medida han tenido influencia en las escuelas. En un intento de hacer las calificaciones lo más precisas y exactas posibles, estos expertos han puesto el énfasis en la importancia de datos de tests y medidas; cuando estos tipos de datos se emplean para asignar notas - afirman - las calificaciones tienden a ser más fiables. Así, en favor de la fiabilidad, se usan todavía las puntuaciones como modos de resumir e informar del logro de los estudiantes (pág. 234). Sin embargo, no debe olvidarse que una puntuación es sólo eso, un punto o estimación puntual del verdadero dominio o rendimiento de un alumno en dicha materia. Por ello, utilizar tal valor numérico (tal punto) como único modo de resumir la información puede ser un error; quizás se gane en fiabilidad, pero se pierde, seguramente, en validez a la hora de interpretar la calificación.

De aquí que coincidamos con otros autores (Beggs y Lewis, Tenbrick, etc.) en que quizás más importante que las puntuaciones (a la hora de informar) sea resumir la información en un informe algo más amplio (no demasiado tampoco) que permita ofrecer una información básica y contextualizada de logros y progresos del alumno. En este sentido conviene diferenciar entre:

-Puntuaciones o datos obtenidos.

-Resumen, modo de averiguar lo que el proceso de evaluación ha producido.

-Informe, modo de comunicar a otros lo que se ha averiguado.

El profesor pues, primero debe hacer un resumen o síntesis, interpretativamente, de los resultados del proceso de evaluación (sobre la base de datos de que disponga), para después elaborar un informe sobre dicho proceso (que normalmente no incluirá todo el resumen de la evaluación).

1. Resumen de los resultados de la evaluación. Este paso, previo a la elaboración de informes, será fácil realizarlo si se siguieron sistemáticamente los 9 pasos anteriores y, más concretamente, si se dispone de un buen registro de información (paso 7). Los pasos fundamentales para elaborar ese resumen son:

-Determinar el objetivo del resumen (sobre quién se hace, para quién se hace, para qué se utilizará).

-Enunciar las decisiones más importantes que se han tomado.

-Indicar los juicios que han influido en cada decisión.

-Resumir la información empleada para formular cada juicio.

2. Informe de los resultados de la evaluación. Los informes pueden ser escritos, orales, formales o informales y pueden realizarse para el conocimiento del profesor o para otras entidades (padres, centro, orientador, etc.).

Dada la diversidad posible de informes distintos que pueden realizarse, no vamos a describirlos uno a uno, baste con enunciar algunos principios básicos que deberán seguirse al elaborarlos, algunas sugerencias para su elaboración y algunas modalidades básicas de informe (Tenbrick, 1981):

-Principios

•Debe ofrecer toda la información necesaria para aquellos a los que se dirige.

•Debe ser claro y fácilmente comprensible para aquellos a los que se dirige.

-Sugerencias para ayudar a la elaboración de un informe.

•Formular claramente aquello de lo que se informa.

•Incluir una guía para interpretar la información que proporcione dicho informe.

•Dar la información tan directa y claramente como sea posible.

•Dar la información en modo que sea significativa para la persona que recibe el informe.

•Cuando sea necesario, explicar cómo la información llevó a los juicios o decisiones que se formularon y tomaron.

Algunos tipos de informe:

•Expediente.

•Cartas-informe.

•Informes diagnóstico.

•Entrevistas con padres.

•Llamadas telefónicas.

En todo caso, el sentido común, la preparación y seguimiento sistemático del proceso evaluativo y el resumen de la evaluación elaborado por el profesor, serán las mejores garantías para que los informes evaluativos sean eficientes y provechosos, tanto para profesores y alumnos, como para los restantes miembros de la comunidad educativa.

TERCERA PARTE

La medida en educación. Teorías de la medida y problemas resueltos

Introducción

La interpretación y uso de las propiedades métricas de los tests e instrumentos de medida precisa ir más allá del cálculo empírico y conocer las bases en que se fundamentan estos cálculos. Ello exige utilizar teorías de la medida, entre las cuales destacan la Teoría clásica de la medida y la Teoría de respuesta al ítem.

Los instrumentos de medida son, seguramente, la tecnología más sofisticada que usamos pedagogos y psicólogos para desarrollar nuestra profesión. Como hemos visto, los tests son muestras de conductas que nos permiten "medir" y llevar a cabo estimaciones e inferencias sobre la conducta de las personas.

Usarlos bien es capital, pues con ello logramos objetivar y estimar variables que nos permiten evaluar rasgos y características de las personas y contribuir así a la mejora y perfeccionamiento de las mismas a través de la educación. Como afirma Muñiz (2010), "no conviene olvidar que los tests nacen con un afán de objetividad y justicia, para evaluar a las personas por lo que realmente valen, evitando evaluaciones sesgadas por aspectos como la cuna, la clase social, la raza, el sexo, las creencias, las cartas de recomendación y otros sistemas de evaluación subjetivos... se trata de evaluar a todos por el mismo rasero".

9

Teorías de la medida

A la elaboración y análisis de los tests subyacen siempre teorías que guían la construcción, que condicionan a los propios instrumentos de medida según los avances teóricos y estadísticos vigentes.

Los tests son, pues, instrumentos complejos para medir y realizar inferencias sobre las personas, que tienen consecuencias en la toma de decisiones sobre las mismas. Por ello, hay que asegurar que estas estimaciones que con los tests realizamos son adecuadas y oportunas, pues de lo contrario podemos perjudicar seriamente a las personas que se "someten" a ellos, al objeto de que podamos ayudarles.

Las teorías estadísticas de los tests van a permitirnos la estimación de las propiedades métricas de los instrumentos de medida, garantizando así su buen uso. Sin esas teorías, no podríamos estimar la fiabilidad y la validez de tales instrumentos, lo cual nos dejaría sin garantías rigurosas y científicas para poder usarlos.

Obviamente, además de estas teorías de la medida, a cada test o instrumento de medida subyace una teoría sustantiva del rasgo, variable o constructo que se mide, teoría que guía siempre la elaboración del instrumento de medida.

Como hemos dicho, dos son los grandes enfoques o teorías de la medida: la Teoría Clásica (TCT) y la Teoría de Respuesta al Ítem (TRI). En este trabajo no expondremos detalladamente estas teorías, sino que simplemente enunciaremos sus conceptos clave, de manera que, como usuarios de los instrumentos de medida tengamos una idea básica y una comprensión suficiente del alcance de las propiedades métricas de los tests o instrumentos de medida que usamos. Para una exposición más detallada de estas teorías, puede acudirse en castellano a las obras de Muñiz (1997, 2000 y 2010) o de Martínez Arias y otros (2005), por ejemplo.

9.1. Teoría clásica de la medida

Es sin duda el enfoque predominante en la construcción y análisis de instrumentos de medida. Como afirma Muñiz (2010) "los diez tests, más utilizados en España, han sido todos ellos, sin excepción, desarrollados bajo la óptica clásica".

Las raíces de la teoría nos remiten, próximamente, a los años finales del siglo xix y principio del siglo xx, con las aportaciones de autores como Galton, Mckeen, Cattell,

Binet y Simón, Terman y Stern, en sus esfuerzos por construir los primeros tests de inteligencia y tareas cognoscitivas. También, los primeros trabajos de Spearman (1901-1920) fueron claves para las primeras formulaciones de la teoría, como Teoría de la puntuación verdadera, los primeros tests de inteligencia que buscaban confrontar su "Teoría de los dos factores de la inteligencia" y el desarrollo de la técnica de análisis factorial, como herramienta metodológica de reducción de datos para la búsqueda de las dimensiones o factores comunes a diversas conductas y tareas intelectuales.

A partir de estos trabajos primeros, otros autores como Holder, Campbell y Stevens posibilitaron el desarrollo de la teoría hasta los años 50 del siglo pasado, que permitieron a Gulliksen (1953) llevar a cabo la primera síntesis formalizada de la Teoría clásica de la medida. Más adelante, Lord y Novick (1968) reformularon la Teoría clásica con un nuevo enfoque, la Teoría de respuesta al ítem, como alternativa a la Teoría clásica de la medida.

Como afirma Muñiz (2010), tras más de cien años en el circuito, se ha ganado por méritos propios el adjetivo de "clásica"... y si nos preguntamos, por curiosidad, cuáles son los tests más utilizados en la actualidad por los profesionales españoles y europeos (pedagogos y psicólogos), la encuesta nos diría que fueron: 16PF, WISC, WAIS, MMPI, Beck, STAI, Rorschach, Raven, Bendov e ISRA, todos ellos desarrollados bajo la Teoría clásica y muchos de ellos a lo largo del siglo xx.

9.1.1. La Teoría de la puntuación verdadera

Cuando nosotros aplicamos un test a un sujeto, obtenemos una puntuación directa (un número), que refleja, supuestamente, un "constructo" subyacente al propio instrumento de medida, que "posee" en cierto grado el sujeto. Un constructo es un atributo que "posee" un sujeto y que supuestamente se refleja en la ejecución de un test", nos dirían Cronbach y Meehl (1956).

La esencia de la medición es, pues, una comparación entre una unidady una cantidad, siendo el resultado de tal comparación un número, nos dirían los viejos textos.

Esta definición anterior es exigente y podría llevarnos a pensar que no es posible realizar tal operación comparativa con conductas humanas. Sin embargo, la verdad es que el concepto de medida admite diferentes grados de experiencia, que sintéticamente puede aceptarse son 4 o 5 reglas o niveles de medición que, como diría Stevens (1950), nos permiten la "asignación" de "numerales" a objetos: nominal, ordinal, de intervalo y de razón (la 5.a podría ser la escala o nivel de medición de "cuasi-intervalos", situada entre los niveles ordinal y de intervalo).

Ciertamente, "una condición imprescindible para poder llevar a cabo la medición es la de Isomorfismo entre operaciones que puede realizarse con los objetos o hechos a los

que alude Stevens y las que permiten llevar a cabo los números. Cuando esto ocurre, podemos asignar números a los objetos o hechos y realizar con tales números las operaciones que son factibles con los objetos o con los hechos; no estará, pues, justificado realizar operaciones o establecer relaciones entre números que no puedan llevar a cabo o establecerse con los objetos empíricos" (Pérez Juste, 1985).

A cada nivel de medición, podemos aplicarle para el análisis de datos la estadística que le corresponda. Al nivel de medición nominal, la estadística de frecuencias, al nivel ordinal, la estadística de rangos, al nivel de medición de cuasi intervalos, intervalos y razones, la estadística más potente, la llamada paramétrica.

Todo instrumento de medida ha de estar validado, homologado; esto es, debe darnos indicaciones del grado de precisión con que medimos (fiabilidad), asegurando además que medimos en grado suficiente lo que queremos o decimos medir (validez).

No olvidemos que en tales instrumentos nos vamos a apoyar para tomar decisiones importantes para las vidas y desarrollo de las personas.

Cuando medimos, debemos estar seguros de que nuestra medida está "garantizada" en un grado suficiente y nuestras puntuaciones directas, obtenidas al aplicar el instrumento, nada nos dice, sobre dichas "garantías".

En la teoría clásica, partimos de una ecuación muy sencilla, formulada en su día por Spearman (1904) en su modelo lineal clásico muy eficaz y parsimonioso.

Toda medida (X) es una estimación de la puntuación verdadera (V). Pero esta estimación conlleva siempre un error (E) que puede ser por sobreestimación o por infraestimación. Se expresa así, formalmente: $X = V \pm E$

El error que se comete puede deberse a muchas razones, que pueden ser internas (de la propia persona) o externas al individuo (contexto, ambiente, etc.). Para avanzar, Spearman formuló tres supuestos al modelo y una definición:

1. La puntuación verdadera (V) es la esperanza matemática de la puntuación empírica: $V = E(X)$, lo que indica que la puntuación verdadera de una persona en un test, es la que se obtendría como media de infinitas aplicaciones del test a esa persona.

2. No hay una relación entre la magnitud de las puntuaciones verdaderas y el tamaño de los errores que afectan a esas puntuaciones: $r(v, e) = 0$.

3. Los errores de medida de la persona en un test no tienen relación con los errores de medida en otro test diferente: $r(ej, ek) = 0$.

Además del modelo, como señala Muñiz (2010), y de estos tres supuestos, se formula

una definición de lo que son tests paralelos, según la cual dos tests distintos pueden medir el mismo rasgo con ítems distintos. Las puntuaciones verdaderas de las personas en distintos tests paralelos serán las mismas y también serían iguales las varianzas de los errores de medida.

9.1.2. Algunas limitaciones de la Teoría clásica de medida

Estos elementos acaban siendo lo central de la Teoría clásica de la medida, que nos lleva a un modelo lineal en donde la Teoría de la puntuación verdadera supone siempre la existencia de errores de medida que pueden estimarse conociendo la fiabilidad del test (r) y sabiendo que los errores de medida se distribuyen normalmente.

Ahora bien, y aunque de la Teoría clásica hay pocas dudas en cuanto a su utilidad y eficacia, pues como se ha dicho, la mayoría de los tests actuales editados en España están desarrollados dentro de este marco, sin embargo presenta algunas limitaciones que hacen que surjan otros modelos de medida, complementarios (Teoría axiomática o Teoría de la generalizabilidad) o alternativos (Teoría de respuesta al ítem).

Las dos cuestiones principales que no encuentran una buena solución en la Teoría clásica son (como señala Muñiz, 2010):

a) Las medidas no resultan invariantes respecto al instrumento o test utilizado. Los tests raramente se miden en la misma escala y la comparación entre puntuaciones se hace dificil, aunque se usan puntuaciones individuales transformadas para tal comparación (Z, Pc, Q, Pentas, Eneatipos, etc.).

   En la práctica, es dificil de garantizar la equiparabilidad de tests distintos.

b) No está asegurada la invarianza de las propiedades de los tests respecto a las muestras usadas para estimarlos (ru, xy, indicadores de los ítems...).

Ambas cuestiones, así como otras menores, encuentran solución en la Teoría de Respuesta al ítem (TRI) de la que hablamos más adelante.

Aún así, conviene aclarar que estos nuevos modelos, como TRI, no invalidan en absoluto la Teoría clásica de la medida, si bien constituyen un complemento muy valioso que en circunstancias particulares solucionan mejor estas cuestiones, no bien resueltas por el enfoque lineal clásico.

9.2. Teoría de respuesta al ítem

La Teoría de Respuesta al Ítem (TRI) surgió de la mano de Lord y Novick (1962/68) al formular una teoría alternativa a la TCT que resolviera algunas cuestiones no bien resueltas, como mencionamos anteriormente.

Para ello, formulan modelos más complejos y menos intuitivos que la TCT, sin que ello suponga dificultades extraordinarias. La investigación actual sobre medida en educación está dominada por el enfoque TRI. Algunos desarrollos informáticos ayudan mucho al desarrollo de los principales modelos de TRI, de uno, dos y tres parámetros. Siguiendo a Muñiz (2000) veamos sus características:

a) Algunas bases:

•La TRI es también denominada Teoría del rasgo latente.

•Intenta dar una fundamentación probabilística a la medición del rasgo y constructos no observables.

•Consideración del ítem como unidad de análisis básica del test, en lugar de las puntuaciones totales (TCT).

•Necesidad de programas informáticos de análisis (BILOG, BIGSTEP, CONQUEST, etc.).

•La respuesta de los sujetos a los ítems depende de:

-El nivel de rasgo (6)

-Los parámetros de los ítems (c)

b) Limitaciones de la TCT

•Los estadísticos de los ítems (dificultad, discriminación, fiabilidad y validez) dependen de la población o muestra donde se calculen.

•Los estadísticos del test (fiabilidad y error típico de medida) dependen de la población o muestra donde se calculen.

•Dificultades derivadas del concepto de paralelismo de medidas. Fiabilidad como correlación entre formas paralelas. Es dificil satisfacer la definición de paralelismo.

•Concepción única e indiferenciada del error. Existen distintas fuentes de error.

•La puntuación en TCT depende del conjunto de ítems que componen el test. No se puede hablar de aptitud en términos generales.

•El estimador de la puntuación verdadera V depende del grupo de referencia. Ecuación de Kelley:

$$p_i X + (1 - p)\ \mu$$

•La TCT no proporciona un modelo teórico para las respuestas de los ítems.

•Escasa plausibilidad de los supuestos sobre el error de medida. Violación del supuesto de homocedasticididad de los errores de medida.

•Medidas sin adecuación óptima a las características de los sujetos. No proporcionan estimaciones precisas para grupos extremos. Se necesitan modelos que permitan emparejar las dificultades de los ítems con los niveles de aptitud del sujeto.

Cuadro 9.1. Diferencias entre TCTy TRI

Características	*TCT*	*TRI*
Modelo	Lineal	No lineal
Nivel de análisis	Test	Ítem
Relaciones ítem/rasgo	No especificadas	Función característica del ítem (CCI)
Supuestos	Débiles y fáciles de cumplir	Fuertes y difíciles de cumplir
Invarianza de estadísticos	No, dependientes de la muestra	Sí, si el modelo ajusta
Estadísticos del ítem	p_i, r_{bp}, r_b	*a, b, c,* según modelos; funciones de información
Tamaño muestral para la calibración	200-500	Depende del modelo y procedimiento de estimación $n > 500$
Estadísticos del test	r_{xt}, S_t	Función de información, error típico

CUADRO 9.2. Cambio de reglas de medida entre TCT y TRI

Reglas	*TCT*	*TRI*
1	Error típico para todas las puntuaciones	El error típico difiere entre puntuaciones o patrones de respuestas
2	Tests largos más fiables que cortos	Aunque la relación entre fiabilidad y longitud se mantiene, tests cortos pueden ser más fiables que largos
3	Comparaciones de puntuaciones sólo cuando las formas son paralelas	Comparación sin necesidad de paralelismo
4	Las características de los ítems dependen de la muestra donde se calculan	Pueden obtenerse estimadores insesgados de los parámetros de los ítems
5	Formatos mixtos de ítem (dificultad) llevan a desequilibrios de ponderación en la puntuación total	Formatos mixtos pueden proporcionar puntuaciones óptimas
6	Test como unidad global de análisis	Propiedades de los ítems
7	Problemas en la medida del cambio	Facilidad para medir el cambio

c) Teoría del rasgo latente (9)

- El rasgo latente es una característica del sujeto que no es directamente observable pero que determina la forma de responder a la prueba que se le presenta (inteligencia, motivación, rendimiento...).

- Un modelo de rasgo latente especifica la relación entre el rendimiento de un sujeto en un determinado test y el rasgo o la habilidad latente.

- Los modelos de rasgo latente establecen una relación matemática entre la cantidad observable y lo inobservable, un modelo matemático.

- El modelo describe las curvas características de los ítems que ponen en relación el rasgo y la probabilidad de respuesta correcta.

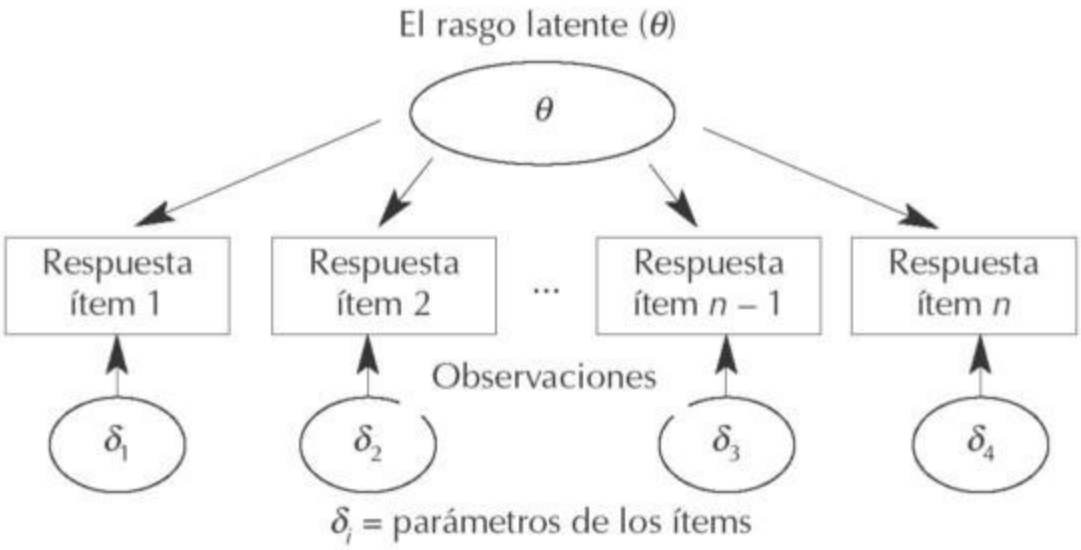


Figura 9.1. Teoría del rasgo latente.

d) Características necesarias de los nuevos modelos psicométricos

•Las características de los ítems no pueden depender del grupo en el que se calculen.

•Las puntuaciones no pueden depender de los contenidos particulares del test.

•El modelo debe expresarse a nivel de ítem.

•No depender del paralelismo u otras formas de equivalencia para calcular la fiabilidad del test.

•Proporcionar medidas de precisión para cada nivel de aptitud, eliminando el supuestode homocedasticidad de la varianza de error.

• Modelos que permitan relacionar el rasgo con el rendimiento del sujeto en el ítem.

e) Características de los modelos TRI

•Suponen la existencia de rasgos o aptitudes latentes que permiten predecir o explicar la conducta de un examinado ante un ítem de un test.

•La relación entre el rasgo y la respuesta del sujeto a un ítem puede describirse mediante una función monótona creciente denominada curva característica del ítem (CCI) o función de respuesta al ítem (FRI).

•La puntuación en el rasgo será la misma con diferentes conjuntos de ítems, aunque el valor del estimador puede variar a causa de los errores de medida y las propiedades

de los ítems.

f) Curva característica del ítem (CCI)

•Caracteriza a cada ítem.

•Describe los cambios en el nivel del rasgo relacionados con cambios en la probabilidad de una respuesta específica.

•Sujetos con diferentes niveles de rasgo tendrán distinta probabilidad de acertar el ítem.

•En ítems dicotómicos predice la probabilidad de éxito del ítem a partir del nivel de rasgo que posee el sujeto.

•En ítems politómicos (p. ej., Lickert, respuesta graduada) predice la probabilidad de las respuestas en cada categoría a partir del nivel de rasgo.

•Es necesario especificar la dimensional ¡dad, es decir, el rasgo o los rasgos que mide un ítem o un test. Unidimensionalidad.

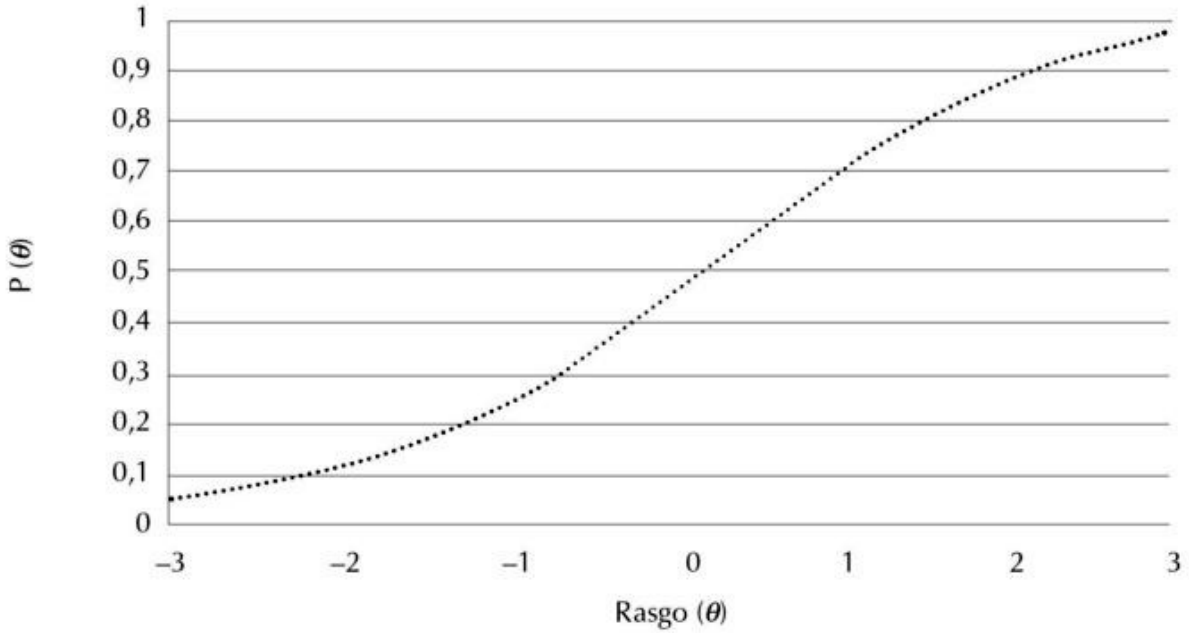


Figura 9.2. Curva característica de un ítem (CCI).

g) Ventajas de los modelos TRI

•Es necesario especificar la dimensionalidad, es decir, el rasgo o rasgos que mide un ítem o un test. Unidimensionalidad.

•Las puntuaciones del sujeto podrán compararse aunque hayan contestado a diferentes conjuntos de ítems.

•Las unidades básicas de análisis son los ítems del test.

•Los estimadores de las propiedades de los ítems no dependen de las muestras de sujetos donde se estiman.

•No se requieren los supuestos de homocedasticidad y paralelismo.

•Tanto el sujeto como el ítem pueden situarse en un punto del espacio del rasgo.

-La diferencia entre habilidad y dificultad (6 - b) la que desempeña un papel fundamental en la probabilidad de la respuesta correcta al ítem.

-Parámetro de dificultad en TRI (b) es la cantidad de rasgo que demanda un ítem para contestarlo de forma correcta. ¿Qué es un TCT?

•Es posible evaluar el ajuste del modelo de forma empírica, es decir, si el modelo se adecua a los datos.

•Invarianza de los parámetros de los ítems. Los parámetros de los ítems no dependen de la muestra utilizada, estarán relacionados linealmente con los obtenidos en otras muestras.

•Invarianza de los parámetros del rasgo. No depende del conjunto de ítems utilizado en la estimación. También habrá una relación lineal entre parámetros estimados con muestras distintas. Permite la comparación de puntuaciones aunque se hayan utilizado ítems distintos.

•Medidas locales de precisión. La TCT utiliza un único valor (la fiabilidad) para evaluar la precisión de todo el conjunto de puntuaciones, en cambio, la TRI lo hace mediante la función de información que indica cómo son de precisas las puntuaciones en diferentes niveles de aptitud.

h) Función de información del test

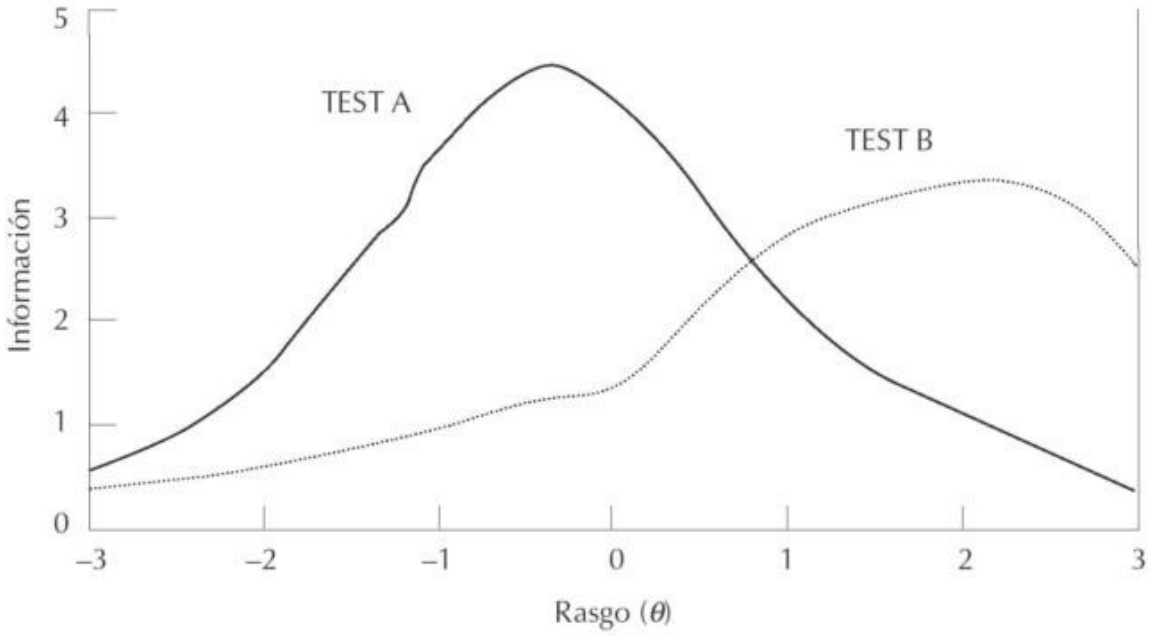


Figura 9.3. Información del test para un determinado nivel de habilidad.

•La información del test para un determinado nivel de habilidad (6) es el resultado de la suma de la información de los ítems que lo componen.

•La información proporcionada por un test depende de la calidad (más discriminación) y el número de ítems que lo forman (la información del test es aditiva).

•Es útil en la construcción de test porque permite evaluar el efecto que produce la introducción o eliminación de ítems en la precisión de las medidas.

•Permite establecer comparaciones entre tests diseñados para evaluar el mismo constructo.

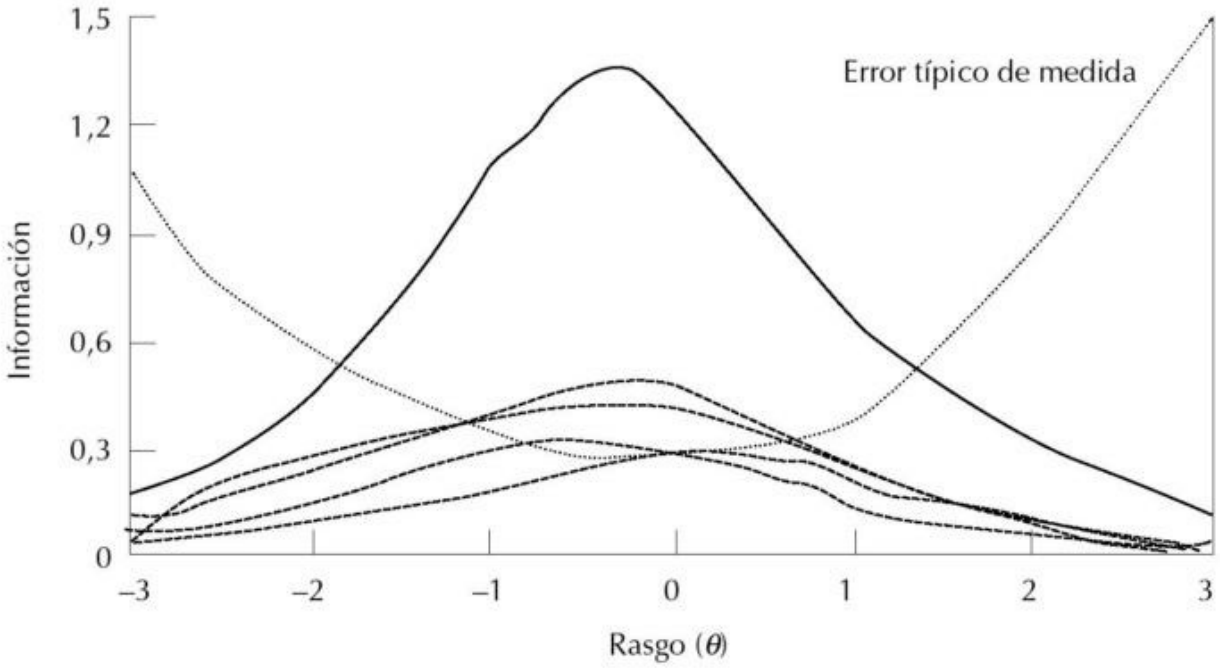


Figura 9.4. Función de información del test: comparaciones entre tests diseñados para evaluar el mismo constructo.

i) Supuestos de la TRI

•Los modelos TRI indican que la probabilidad de responder a un ítem correctamente depende de la aptitud (rasgo) del sujeto y de las características del ítem.

•Los modelos incluyen supuestos acerca de los datos, que pueden probarse aplicando pruebas. Son dos principalmente:

-Unidimensionalidad: la TRI asume que hay un conjunto de rasgos responsables de la actuación del sujeto en un test. En la práctica se supone que una única aptitud o rasgo es suficiente para explicar los resultados (unidimensionalidad). Normalmente, aunque diferentes factores pueden afectar en el momento de la prueba (atención, motivación, ansiedad, etc.), se habla de un rasgo dominante que explica las respuestas. En caso de asumir más de un rasgo, es necesario un modelo multidimensional (en desarrollo).

-Independencia local: la probabilidad de responder correctamente a un ítem es independiente de la probabilidad de responder correctamente a cualquier otro.

10

Resolución de problemas sobre medición en educación

La medida es un elemento imprescindible de la investigación y evaluación en el campo de las Ciencias Sociales. Sin embargo, en este área plantea muchos problemas por la naturaleza intrínseca del objeto a medir. La propia definición de las variables y, en consecuencia, la aproximación a su medida son sus principales dificultades. De aquí el interés que para el investigador y evaluador debe tener la construcción de instrumentos técnicamente elaborados que reúnan las características mínimas exigidas a un buen instrumento de medición.

Aunque no se puede decir que la medida sea propiamente estadística, sin embargo, la utiliza para el estudio técnico de sus instrumentos. Por esto, y dada la importancia ya subrayada en nuestra área, con este capítulo y el anterior se facilita la comprensión del estudio de las características técnicas de los instrumentos de medición. Nos hemos centrado esencialmente en el contexto de la teoría clásica de los tests, ya que, por el momento y en nuestro país, los tests comercializados y de uso más frecuente están elaborados según estos planteamientos. No obstante, incluimos también algunos ejercicios referidos a tests elaborados según la Teoría del rasgo latente o Teoría de Respuesta al ítem (TRI), una alternativa de gran actualidad e interés. Estos ejercicios son fundamentalmente de interpretación de datos o gráficas, ya que la complejidad del análisis hace que se realice a través de programas de ordenador.

Dentro del contexto de la teoría clásica, nos ocupamos de la fiabilidad, validez, análisis de ítems y tipificación. La ejemplificación del estudio de estas características se realiza a partir de una prueba o instrumento de medida, compuesto de 50 ítems y aplicado a 46 sujetos. Aunque, lógicamente, el tamaño de las muestras de sujetos para estos estudios suele ser bastante mayor, hemos decidido reducir el número para facilitar los cálculos. Los resultados en la prueba se presentan al iniciar el apartado de ejercicios dedicados a fiabilidad. Concedemos especial importancia a la valoración de los resultados. Somos conscientes de las limitaciones que supone referirnos únicamente a un tipo de prueba con características específicas en cuanto a la valoración de los ítems, la naturaleza de las variables, etc. Esto nos lleva al uso de unas determinadas técnicas de análisis, no permitiendo la práctica de otras. No obstante, consideramos que el ejemplo resulta bastante completo y clarificador. Las posibles variaciones vendrían dadas por el uso de estadísticos, ya estudiados en anteriores capítulos como, por ejemplo, las

correlaciones. También se incluyen algunos ejercicios complementarios al análisis del instrumento, como práctica de la relación entre algunas características técnicas con otras variables (longitud, variabilidad, etc.). (La mayoría de los problemas y fórmulas están tomados de Fdez. Díaz, M.J.; García Ramos, J.M.; Fuentes, A. y Asensio, l., 1990 y Fdez. Díaz, M.J.; García Ramos, J.M.; Asensio, I. y Fuentes, A. 1990).

io.i. Teoría clásica de la medida

Fiabilidad

- FIABILIDAD
  - A) *FIABILIDAD RELATIVA:* Proced. de estimación del coefic. de fiabilidad
    - 1. Como estabilidad: Aplicación de una misma prueba en dos ocasiones sucesivas: coeficiente de correlación entre las dos series de puntuaciones.
    - 2. Como equivalencia: Aplicación de dos formas paralelas de un instrumento de medida: coeficiente de correlación entre las puntuaciones en las dos formas paralelas.
    - 3. Como consistencia interna: Se aplica la prueba una sola vez
      - 3.1. Proced. de las mitades (división de los elementos en dos mitades diferentes, ítems pares e impares):
        - 3.1.1. Spearman-Brown.
        - 3.1.2. Rulon
        - 3.1.3. Guttman
      - 3.2. Proced. De Kuder-Ricardson
      - 3.3. α de Cronbach
  - B) *FIABILIDAD ABSOLUTA*
    - Fiabilidad en términos de varianza: $r_{xx} = \frac{S_v^2}{S_t^2} = \frac{S_t^2 - S_e^2}{S_t^2}$
    - Error típico de medida: $\sigma_{med.}$
    - Error típico de diferencia de medidas: $\sigma_{dif.\ med.}$
  - C) *RELACIÓN DE LA FIABILIDAD CON OTRAS VARIABLES*
    - 1. Fiabilidad y longitud.
    - 2. Fiabilidad y variabilidad.

Figura 10.1. Fiabilidad. Tipos y aplicaciones.

io.i.i. Formulario

A) Coeficiente de fiabilidad:

A.1) Como estabilidad-co eficiente de correlación.

A.2.) Como equivalencia: coeficiente de correlación.

A.3.) Como consistencia interna:

A.3.1.) Procedimiento de Spearman-Brown:

-Cálculo de la correlación entre las dos mitades.

-Estimación del coeficiente de fiabilidad (R) para el test completo:

$$R_{xx} = \frac{n \cdot r_{xx}}{1 + (n-1) r_{xx}} = \frac{2 r_{xx}}{1 + r_{xx}}$$

-

n = número de veces que el test se alarga o acorta.

$$n = \frac{\text{número de ítems que va a tener la prueba}}{\text{número de ítems que tiene la prueba}}$$

A.3.2.) Procedimiento de Rulon:

$$r_{xx} = 1 - \frac{S_d^2}{S_t^2}$$

Sd = varianza de las diferencias de puntuaciones.

S2=varianzat total.

A.3.3.) Procedimiento de Guttman:

$$r_{xx} = 2\left(1 - \frac{S_1^2 a + S_2^2 a}{S_t^2}\right)$$

S1 S. a = varianza de la primera mitad.

Sz S. a = varianza de la segunda mitad.

A.3.4.) Procedimiento de Kuder-Richardson:

$$r_{xx} = \left(\frac{n_e}{n_e - 1}\right) \left(\frac{S_t^2 - \Sigma pq}{S_t^2}\right) \rightarrow \text{ para ítems dicotómicos.}$$

p = proporción de aciertos.

q= 1-p

n=e número de elementos.

$$r_{xx} = \frac{n_e S_t^2 - \bar{X}(n_e - \bar{X})}{S_t^2 (n_e - 1)} \rightarrow \text{ para elementos de dificultad similar.}$$

$$r_{xx} = \frac{n_e}{n_e - 1} \cdot \frac{S_t^2 - \Sigma S_e^2}{S_t^{22}} \rightarrow \text{ para elementos no dicotómicos.}$$

S2 = varianza de cada elemento. e

A.3.5.) a de Cronbach:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma_{i=1}^{k} S_i^2}{S_t^2}\right)$$

k = número de ítems del test.

S'=t varianza de cada ítem o subtest.

~ = varianza de la suma de todos los ítems o subtests que componen el test.

$$\alpha_S = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \rightarrow$$ = fórmula para utilizar cuando las observaciones de cada ítem están estandarizadas.

r = correlación promedio entre todos los ítems del test.

B) Fiabilidad absoluta

- $r_{xx} = \frac{S_v^2}{S_t^2} = \frac{S_t^2 - S_e^2}{S_t^2} = 1 - \frac{S_e^2}{S_t^2}$

S2=V varianza verdadera.

S2=e varianza de error.

S2=t varianza total.

•Error típico de medida: a1ea:

$$\sigma_{\text{med.}} = S_x \sqrt{1 - r_{xx}}$$

Sx= desviación típica del test.

•Error típico de diferencia de medidas:

$$\sigma_{\text{dif. med.}} = S_x \sqrt{1 - r_{xx}} \sqrt{2} \qquad \text{(dentro del mismo test)}$$

$$\sigma_{\text{dif. med.}} = S_x \sqrt{1 - r_{AA} - r_{BB}} \qquad \text{(en tests diferentes, A y B)}$$

•Aplicación de los errores típicos:

-Significación de diferencias:

$$RC = Z = \frac{\text{Diferencia empírica}}{\text{Error típico de diferencias}}$$

Si la diferencia empírica entre las puntuaciones se presenta en unidades típicas, el valor de S que aparece en las fórmulas de los errores típicos será lógicamente igual a 1.

Estimación de la puntuación verdadera de un sujeto:

$$IC = X_i + \sigma_{\text{med.}} Z_{nc}$$

C) Relación de la fiabilidad con otras variables:

1. Fiabilidad y longitud:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$$

$$n = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

2. Fiabilidad y variabilidad:

$$S_A\sqrt{1-r_{AA}} = S_B\sqrt{1-r_{BB}}$$

10.1.2. Ejercicios resueltos

Para la realización de este ejercicio práctico sobre el estudio de la fiabilidad de un instrumento de medida, dentro del contexto de la teoría clásica de los tests, vamos a partir de los datos del cuadro 10.1. En él se recogen los resultados obtenidos por 46 sujetos en una supuesta prueba o instrumento de medida compuesto de 50 ítems. Los signos positivos indican acierto del ítem, mientras que los errores se presentan en blanco. No hay ningún ítem sin responder por parte de ningún sujeto. Asimismo, utilizaremos este mismo cuadro para la realización de los ejercicios prácticos del estudio de la validez del análisis de ítems y para la tipificación o baremación de resultados. También plantearemos problemas relacionados con las características mencionadas, cuyos cálculos se apoyarán en los mismos datos.

Se presentan también resultados parciales al margen del cuadro 10.1 para realizar los cálculos que se precisen.

1. A partir de los datos del cuadro 10.1 estimar la fiabilidad de la prueba

SOLUCIÓN:

Para estimar la fiabilidad en este caso, deberemos utilizar alguno de los procedimientos de consistencia interna, ya que no disponemos de resultados de la aplicación de la prueba en una segunda ocasión ni los resultados en una prueba paralela.

La solución en estos dos últimos casos resultaría únicamente del cálculo de la correlación entre las dos series de puntuaciones, por lo que técnicamente el cálculo es muy sencillo, pero su uso no es muy frecuente por las dificultades de otro orden que plantean.

Así pues, para la obtención de la fiabilidad utilizaremos en la práctica los procedimientos de Spearman-Brown, Rulon, Guttman, Kuder-Richardson y a de Cronbach. El cálculo de este último coeficiente, de uso muy frecuente, resulta,

generalmente, muy largo cuando se realiza de forma manual, por lo que se pueden utilizar programas de ordenador al respecto. No obstante, en este caso, dado que los ítems son dicotómicos, el cálculo se simplifica, como veremos posteriormente.

La valoración de los ítems es de 1 (acierto: +) y 0 (error: -). No hubo ningún ítem sin contestar.

a) Procedimiento de Spearman-Brown:

La utilización de este procedimiento exige homogeneidad de las varianzas correspondientes a las dos mitades, por lo que primeramente comprobaremos este supuesto calculando, en primer lugar, ambas varianzas (SI • a y SI • b). Para ello, partimos de la siguiente fórmula:

$$S^2 = \frac{\Sigma_{i=1}^{N} X_i^2}{N} - \left(\frac{\Sigma_{i=1}^{N} X_i}{N} \right)^2$$

CUADRO 10.1. *Puntuaciones obtenidas por cada sujeto en cada ítem*

Sujetos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	Puntuación Total X	p	q
Ítems	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	219	p = proporción de aciertos	q = 1 − p
1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,60	0,77	0,23
2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,72	0,28	0,28
3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,54	0,46	0,25
4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,61	0,39	0,24
5	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,74	0,26	0,19
6													+																												+						0,13	0,87	0,11
7	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,59	0,41	0,24
8				+						+	+																																				0,25	0,85	0,21
9	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,59	0,41	0,24
10	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,78	0,22	0,16
11	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,60	0,37	0,23
12	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,67	0,33	0,22
13		+		+	+	+							+		+													+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,61	0,39	0,24
14	+	+		+	+	+							+		+																																0,35	0,65	0,23
15	+	+		+	+	+							+		+																																0,56	0,44	0,24
16	+	+													+	+			+			+					+		+	+	+	+															0,35	0,65	0,23
17		+		+					+		+																	+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	0,74	0,26	0,19
18	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,74	0,26	0,19
19	+	+	+	+	+		+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,59	0,41	0,24
20	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,81	0,19	0,15
21		+			+				+		+				+						+																										0,35	0,65	0,22
22	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,81	0,22	0,17
23					+								+		+								+		+					+	+															0,17	0,83	0,14
24	+	+	+	+	+		+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,70	0,30	0,21
25								+					+	+	+											+	+		+	+	+	+										+	+			0,19	0,74	0,14
26		+	+		+					+	+				+									+	+	+	+																				0,22	0,67	0,15
27																																															0,04	0,96	0,04
28										+					+												+		+																		0,10	0,89	0,11
29		+		+													+											+	+																		0,11	0,87	0,11
30		+											+		+												+		+	+	+	+															0,61	0,39	0,24
31		+							+						+		+					+	+	+			+		+	+	+	+										+					0,74	0,26	0,19
32		+		+							+	+	+		+		+					+	+	+			+		+	+												+			+	+	0,75	0,25	0,19
33		+				+			+				+		+		+					+	+	+			+		+	+	+	+															0,75	0,25	0,19
34		+		+		+			+	+			+		+		+					+	+	+			+		+	+	+	+															0,75	0,25	0,19
35		+		+		+			+	+					+		+					+	+	+			+		+	+	+	+															0,69	0,31	0,24
36		+		+		+			+	+			+		+		+					+	+	+			+		+	+	+	+															0,69	0,31	0,24
37		+																																			+										0,13	0,85	0,11
38		+			+	+			+				+				+					+		+			+		+	+	+	+															0,30	0,70	0,21
39		+		+		+			+						+									+																							0,40	0,60	0,24
40		+	+	+	+	+		+	+		+	+			+		+	+	+	+		+		+	+	+	+	+	+	+	+	+										+	+	+	+		0,61	0,39	0,24
41	+			+	+	+			+		+		+		+				+	+	+	+	+	+			+		+	+	+	+					+										0,75	0,25	0,19
42	+			+	+				+		+	+	+				+		+	+				+	+	+	+			+		+		+	+		+								+	+	0,75	0,25	0,19
43	+	+							+	+	+	+			+				+	+		+		+	+	+	+	+	+	+	+	+														+	0,75	0,25	0,19
44	+	+	+	+	+				+	+	+	+	+		+	+			+	+		+	+	+	+	+	+	+	+	+	+	+															0,70	0,30	0,21
45		+		+	+	+	+		+	+	+	+	+		+		+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0,87	0,13	0,22
46	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1,00	0,00	0,00
ΣX^2																																																	

Tal como se puede observar:

$$\Sigma X_p = 558 \qquad \Sigma X_p^2 = 8.078 \qquad N = 46$$

$$\Sigma X_i^2 = 521 \qquad \Sigma X_i^2 = 7.435$$

Sustituyendo en la fórmula anterior:

$$S^2_{p(1.^a \text{ mitad})} = \frac{8.078}{46} - \left(\frac{558}{46}\right)^2 = 175{,}60 - 147{,}14 = 28{,}46$$

$$S^2_{i(2.^a \text{ mitad})} = \frac{7.435}{46} - \left(\frac{521}{46}\right)^2 = 161{,}63 - 128{,}28 = 33{,}35$$

CUADRO 10.2. Valores parciales obtenidos a partir del cuadro de resultados correspondientes a las respuestas dadas por el grupo de sujetos

Sujetos	X_p	X_p^2	X_i	X_i^2	$X_p \cdot X_i$	$\frac{(X_p - X_i)}{d}$	d^2	X_T	X_T^2
1	13	169	8	64	104	5	25	21	441
2	16	256	20	400	320	–4	16	36	1296
3	9	81	7	49	63	2	4	16	256
4	18	324	16	256	324	2	4	34	1156
5	12	144	11	121	132	1	1	23	529
6	13	169	12	144	156	–1	1	25	625
7	3	9	4	16	12	–1	1	7	49
8	3	9	2	4	6	1	1	5	25
9	16	256	17	289	272	–1	1	33	1089
10	15	225	7	49	105	8	64	22	484
11	14	196	9	81	126	5	25	23	529
12	13	169	7	49	91	6	36	20	400
13	15	225	13	169	195	2	4	28	784
14	7	49	2	4	14	5	25	9	81
15	21	441	20	400	420	1	1	41	1681

Sujetos	X_p	X_p^2	X_i	X_i^2	$X_p \cdot X_i$	$\frac{(X_p - X_i)}{d}$	d^2	X_T	X_T^2
16	7	49	5	25	35	2	4	12	144
17	13	169	11	121	143	2	4	24	576
18	3	9	4	16	12	–1	1	7	49
19	16	256	13	169	208	3	9	29	841
20	17	289	14	196	238	3	9	31	961
21	14	196	14	196	196	0	0	28	784
22	3	9	4	16	12	–1	1	7	49
23	20	400	17	289	340	3	9	37	1369
24	10	100	11	121	110	–1	1	21	441
25	18	324	23	529	414	–5	25	41	1681
26	13	169	15	225	195	–2	4	28	784
27	15	225	16	256	240	–1	1	31	961
28	17	289	23	529	391	–6	36	40	1600
29	17	289	19	361	323	–2	4	36	1296
30	12	144	11	121	132	1	1	23	529
31	7	49	7	49	49	0	0	14	196
32	9	81	11	121	99	–2	4	20	400
33	6	36	13	169	78	–7	49	19	361
34	8	64	7	49	56	1	1	15	225
35	20	400	19	361	380	1	1	39	1521
36	4	16	4	16	16	0	0	8	64
37	14	196	11	121	154	3	9	25	625
38	16	256	13	169	208	3	9	29	841
39	19	361	16	256	304	3	9	35	1225
40	13	169	11	121	143	2	4	24	576
41	22	484	22	484	484	0	0	44	1936
42	4	16	4	16	16	0	0	8	64
43	5	25	3	9	15	2	4	8	64
44	9	81	7	49	63	2	4	16	256
45	6	36	6	36	36	0	0	12	144
46	13	169	12	144	156	1	1	25	625
Σ	558	8078	521	7435	7586	35	413	1079	30613

X_p = puntuaciones obtenidas por cada sujeto en los ítems pares (primera mitad).
X_i = puntuaciones obtenidas por cada sujeto en los ítems impares (segunda mitad).
X_t = puntuaciones obtenidas por cada sujeto en el total de ítems.

Para el contraste de homogeneidad de varianzas utilizaremos la prueba de Hartley.

$$S^2_{p(1\text{ ª mitad})} = 28,46$$

$$S^2_{i(2.\text{ª mitad})} = 33,35$$

$$F_{\text{máx.}} = \frac{S^2_{\text{mayor}}}{S^2_{\text{menor}}} = \frac{33,35}{28,46} = 1,17$$

Comparación con el valor de Fiteo para grados de libertad: K= 2 (número de muestras), gl=n-1=46-1=45, y a= 0,01:

$$F_{\text{máx crítico}} \approx 2,29$$

Como el valor de F=1,17empírico es menor que el crítico, aceptamos hipótesis nula y, por tanto, podemos afirmar que no existen diferencias significativas entre las varianzas a un a= 0,01.

Así pues, podemos utilizar el procedimiento de Spearman-Brown para el cálculo del coeficiente de fiabilidad. Aplicaremos la fórmula correspondiente:

$$R_{xx} = \frac{2r_{xx}}{1 + r_{xx}}$$

Como desconocemos el valor de r, deberemos obtenerlo a través de la correlación entre las puntuaciones en los ítems pares e impares. Utilizaremos el coeficiente de correlación de Pearson por el procedimiento de puntuaciones directas:

$$r_{xy} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[N\Sigma X_2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$r_{xy} = \frac{(46)\cdot(7.590) - (558)\cdot(521)}{\sqrt{[(46)\cdot(8.078) - (558)^2][(46)\cdot(7.435) - (521)^2]}} = 0,896$$

es el valor del coeficiente de correlación entre los ítems pares e impares, es decir, rr..

Sustituyendo en la fórmula anterior, tendremos:

$$R_{xx} = \frac{2\cdot 0,896}{1 + 0,896} = 0,945 \text{ valor del coeficiente de fiabilidad.}$$

b) Procedimiento de Rulon:

Fórmula:

$$r_{xx} = 1 - \frac{S_d^2}{S_t^2}$$

Sd = varianza de las diferencias de las puntuaciones.

Para calcular la SS calculamos las diferencias entre las puntuaciones de los ítems pares e impares, y las elevamos al cuadrado. Los sumatorios de estos resultados son:

$\sum X_{\text{dif.}} = 35$ $\qquad$ $\sum X_{\text{dif.}}^2 = 413$ (véase cuadro anterior)

$$S_{\text{dif.}}^2 = \frac{413}{46} - \left(\frac{35}{46}\right)^2 = 8{,}4$$

Ahora S-

$$S_t^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} = \left(\frac{\sum_{i=1}^{N} X_i}{N}\right)^2 \qquad \begin{array}{l} \sum X_i = 1.079 \\ \sum X_i^2 = 30.613 \end{array}$$

$$S_t^2 = \frac{30.613}{46} = \left(\frac{1.079}{46}\right)^2 = 115{,}3$$

e z

Sustituyendo en la fórmula de Rulon:

$$r_{xx} = 1 - \frac{8{,}4}{115{,}3} = 1 - 0{,}072 = 0{,}928$$

c) Procedimiento de Guttman:

Fórmula:

$$r_{xx} = 2\left(1 - \frac{S_{1^{\text{a}}}^2 + S_{2^{\text{a}}}^2}{S_t^2}\right)$$

En el apartado a) de este problema calculamos las varianzas correspondientes a la primera y a la segunda mitad o a los ítems pares e impares, y en el apartado b) la varianza total. Así pues, sustituiremos los valores obtenidos en la fórmula:

$$S^2_{1.^a} = 28,46$$

$$S^2_{2.^a} = 33,35$$

$$S^2_t = 115,3$$

$$r_{xx} = 2\left(1 - \frac{28,46 + 33,35}{115,3}\right) = 0,928$$

Como podemos observar la fiabilidad obtenida por este procedimiento coincide con la obtenida por el de Rulon y es prácticamente igual que la obtenida por el procedimiento de Spearman-Brown. En todos los casos, es una fiabilidad alta.

d) Procedimiento de Kuder-Richardson:

Como los ítems son dicotómicos, utilizaremos la fórmula correspondiente para este caso:

$$r_{xx} = \left(\frac{n_e}{n_e - 1}\right) \left(\frac{S^2_t - \sum pq}{S^2_t}\right)$$

pq=10,09 (véanse marginales cuadro de datos)

$$r_{xx} = \frac{50}{50 - 1}\left(\frac{115,3 - 10,09}{115,3}\right) = 1,02 \times 0,912 = 0,930$$

coeficiente de fiabilidad, cuyo resultado coincide prácticamente con los obtenidos anteriormente.

e) Procedimiento del (x de Cronbach:

En este caso, el procedimiento de cálculo del ade Cronbach se simplifica, ya que la varianza de cada ítem es igual ap x q, por ser los ítems dicotómicos. Si los ítems no fueran así, habría que calcular la varianza de cada uno de los ítems o la correlación media entre todos los ítems del test.

En este caso, la fórmula del a de Cronbach es:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} S_i^2}{S_t^2}\right)$$

coincide con la de Kuder-Richardson, por lo que el resultado será también idéntico.

Así, teniendo en cuenta los datos anteriores, tenemos:

$$S_t^2 = 115{,}3$$

$$\sum_{i=1}^{k} S_i^2 = \sum pq = 10{,}09$$

$$k = n_e = \text{número de elementos} = 50$$

Sustituyendo:

$$\alpha = \frac{50}{50-1}\left(1 - \frac{10{,}09}{115{,}3}\right) = 0{,}93$$ valor igual que el obtenido por el procedimiento anterior.

2. Teniendo en cuenta que la fiabilidad del instrumento de medida objeto de estudio es de 0,93, según los resultados obtenidos en el ejercicio anterior:

a) ¿Cuál es la proporción de varianza de error del test?

b) ¿Cuál sería la fiabilidad si se añadieran al test 8 ítems?

c) ¿Cuántos elementos habría que eliminar del test si se desea que la fiabilidad sea de 0,90?

d) Suponiendo que un sujeto hubiera obtenido en el test una puntuación de 30 puntos, ¿entre qué límites se encontraría la verdadera puntuación de éste a un nivel de confianza del 99%?

e) Sabiendo que dos sujetos han obtenido en este test sendas puntuaciones de X~ = 16y X1 = 20, ¿existen diferencias significativas entre ambas puntuaciones? (c = 0, 02)

f) Suponiendo que un sujeto hubiera obtenido en este test una puntuación X, = 40y en otra prueba (B) diferente, una puntuación de 18 puntos, ¿existen diferencias significativas entre ambos resultados, sabiendo que en esta prueba B la fiabilidad es de 0, 85, la media aritmética de 10 puntos y la desviación típica de 5? (a = 0,01).

g) Estimar la fiabilidad del mismo test si se aplica a una muestra de 60 sujetos, sabiendo

que la desviación típica de este grupo es igual a 10.

SOLUCIÓN:

a)Para la resolución de esta cuestión, aplicaremos la fórmula de la fiabilidad absoluta, partiendo del concepto de la fiabilidad como proporción de la varianza verdadera, que es de la varianza total. Como lo que se nos pide es la proporción de varianza de error sobre la varianza total deberemos restar de 1 la proporción de varianza verdadera, según podemos observar en la siguiente fórmula:

$$r_{xx} = \frac{S_v^2}{S_t^2} = \frac{S_t^2 - S_e^2}{S_t^2} = 1 - \frac{S_e^2}{S_t^2}$$

$$\frac{S_e^2}{S_t^2} = 1 - r_{xx};\ \frac{S_e^2}{S_t^2} = 1 - 0{,}93 = 0{,}17 \text{ proporción de varianza error.}$$

b)Aplicaremos la fórmula que relaciona fiabilidad y longitud:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}};\ n = \frac{\text{número de ítems que va a tener el test}}{\text{número de ítems que tiene el test}}$$

$$n = \frac{50+8}{50} = 1{,}16;\ r_{xx} = 0{,}938$$

$$R_{xx} = \frac{1{,}16 \times 0{,}93}{1+(1{,}16-1)0{,}93} = 0{,}938 \text{ nuevo coeficiente de fiabilidad.}$$

c)Aplicamos la misma fórmula que para la cuestión anterior:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$$

y despejamos n:

$$n = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

•La fiabilidad que se quiere alcanzar es de 0,90.

•La fiabilidad de la prueba (r) es 0,93.

•Calculamos n, sustituyendo los valores:

$$n = \frac{0,90(1-0,93)}{0,93(1-0,90)} = 0,677$$

Sabiendo que:

$$n = \frac{\text{número de ítems que va a tener el test}}{\text{número de ítems que tiene el test}}$$

sustituimos los valores para conocer el número de ítems que va a tener el test:

$$0,677 = \frac{\text{número de ítems que va a tener el test}}{50}$$

Número de ítems que tendrá el test = 50 x 0,677 = 33,85 ~ 34

50 - 34 = 16 ítems habrá que eliminar del test para que su fiabilidad sea de 0,90.

d)Para la estimación del intervalo de confianza necesitamos calcular el error típico de medida. Así:

$$IC = X_i \pm \sigma_{\text{med.}} Z_{nc}$$

$$\sigma_{\text{med.}} = S_x \sqrt{1 - r_{xx}}$$

$$X_i = 30$$

$$Z_{nc} = 95\% = 2,58$$

$$S_x = \text{desviación típica del test.}$$

Como obtuvimos la varianza total en el apartado b) del ejercicio anterior, podemos calcular la S:x

$$S_x = \sqrt{115,3} = 10,74$$

$$\sigma_{med.} = 10,74\sqrt{1-0,93} = 2,79$$

$$IC = 30 \pm 2,79 \cdot 2,58 = 30 \pm 7,2 = \begin{cases} 37,2 \\ 22,8 \end{cases}$$

Entre 22,8 y 37,2 se encontrará la verdadera puntuación del sujeto a un nivel de confianza del 99%.

e)Para conocer si existen diferencias significativas entre las puntuaciones obtenidas en la misma prueba, aplicaremos la razón crítica:

$$RC = Z = \frac{\text{Diferencia empírica}}{\text{Error típico de diferencias}}$$

El error típico de diferencias de medias se calculará teniendo en cuenta que es en el mismo test:

$$\sigma_{\text{dif. med.}} = S_x\sqrt{1 - r_{xx}}\sqrt{2}; S_x = 10{,}74 \text{ (obtenido en el apartado anterior)}$$

$$\sigma_{\text{dif. med.}} = 10{,}74\sqrt{1 - 0{,}93}\sqrt{2} = 3{,}93$$

$$RC = 3{,}93$$

$$RC = \frac{|16 - 20|}{3{,}93} = 1{,}017$$

Valor crítico de Za un a = 0,02 -* 2,33 (véanse tablas de curva normal)

1,0 17 < 2,33: no existen diferencias significativas entre ambas puntuaciones. La diferencia no es real, sino aleatoria.

f)Para conocer si existen diferencias significativas entre las puntuaciones obtenidas en dos pruebas diferentes, deberemos proceder como en el caso anterior, si bien variará la obtención del error típico de diferencia de medidas, ya que habrá que tener en cuenta las características de la prueba B(B(X)=10, S=5,rBB=0,85). Así pues:

$$RC = \frac{\text{Diferencia empírica}}{\sigma_{\text{dif. med.}}}$$

$$\sigma_{\text{dif. med.}} = S_x\sqrt{2 - r_{AA} - r_{BB}}; S_x = 1, \text{en unidades típicas}$$

$$\sigma_{\text{dif. med.}} = S_x\sqrt{2 - r_{AA} - r_{BB}} = \sqrt{2 - 0{,}93 - 0{,}85} = 0{,}47$$

Como el error típico se ha calculado para su aplicación en unidades típicas y, dado que los dos tests tienen distinta escala, expresaremos ambas puntuaciones directas en puntuaciones típicas:

$$Z_i = \frac{40 - 23,456}{10,74} = 1,54 \text{ test 1.º:}$$

$$X_i = 40$$

$$\bar{X} = \frac{\sum X_i}{N} = \frac{1.079}{46} = 23,456$$

$S_x = 10,74$ (obtenido en el apartado *b)* del ejercicio anterior)

$$Z_i = \frac{18 - 10}{5} = 1,6 \text{ test B:}$$

$X_i = 18$

$\bar{X} = 10$

$S_x = 5$

Aplicamos la fórmula de la RC:

$$RC = \frac{|1,54 - 1,6|}{0,47} = 0,127$$

Valor crítico de Za un a = 0,01 - 2,58

0,127 < 2,58: no existen diferencias significativas entre las puntuaciones obtenidas en los dos tests a un nivel de significación del 0,01. La diferencia es aleatoria.

g)Para la resolución de este ejercicio aplicaremos la fórmula que relaciona fiabilidad y variabilidad:

$$S_A = \sqrt{1 - r_{AA}} = S_B\sqrt{1 - r_{BB}}$$

Sabemos que:

S_A= desviación típica del test en la muestra de 46 sujetos = 10,737.

$S_A = \sqrt{115,3} = 10,737.$

r_{AA} = 0,93

S_B= desviación típica del test en la muestra de 30 sujetos = 10.

r_{BB}= fiabilidad del test en la nueva muestra.

Sustituyendo:

$$10,737 = \sqrt{1 - 0,93} = 10\sqrt{1 - r_{BB}}$$

$$0,75 = 10\sqrt{1 - r_{BB}}$$

elevamos ambos términos al cuadrado:

$$(0,75)^2 = (10\sqrt{1 - r_{BB}})^2$$
$$0,5625 = 100(1 - r_{BB})$$
$$0,5625 = 100 - 100r_{BB}$$
$$r_{BB} = \frac{100 - 0,5625}{100} = 0,994 \text{ fiabilidad del test en la muestra de 60 sujetos y desviación típica de 10.}$$

10.1.3. Validez

Aunque ofrecemos un esquema general de los distintos tipos de validez, a efectos de cálculo nos ocuparemos de la validez relacionada con un criterio externo (concurrente y predictiva), ya que la validez de contenido y aparente no precisan, como sabemos, cálculo estadístico, y la validez de constructo exige el uso de técnicas estadísticas multivariadas complejas que no son objeto de nuestro trabajo y que, por otra parte, suelen repetir la acumulación de resultados de distintas metodologías.

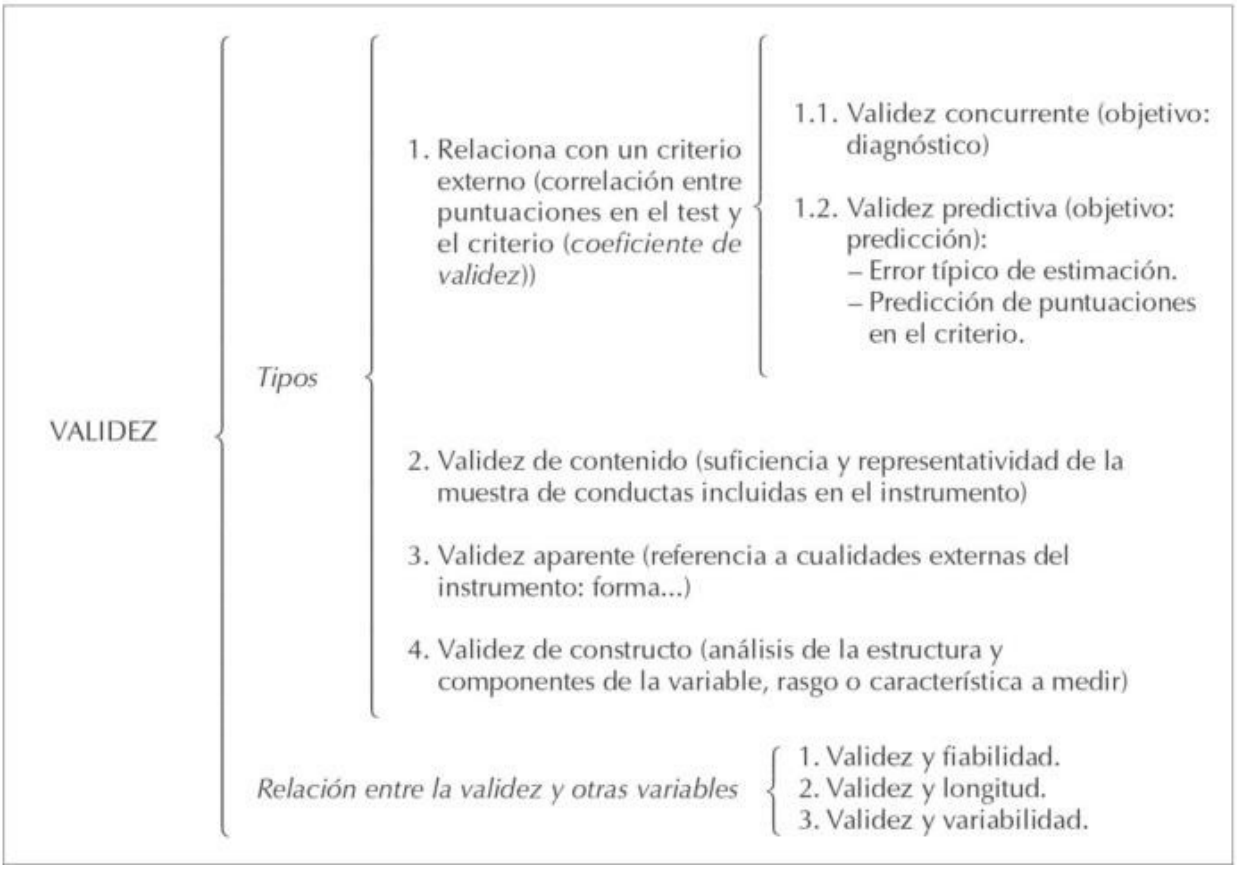


Figura 10.2. Validez. Tipos y aplicaciones.

10.1.3.1. Formulario

A) Coeficiente de validez:

-Coeficiente de correlación entre puntuaciones en el instrumento de medida a validar y el criterio.

B) Predicción de puntuaciones en el criterio:

-Ecuaciones de regresión:

•En puntuaciones directas:

$$Y' = r_{xy}\frac{S_y}{S_x}X_i - r_{xy}\frac{S_y}{S_x}\bar{X} + \bar{Y}$$

$$Y' = r_{xy}\frac{S_y}{S_x}(X_i - \bar{X}) + \bar{Y}$$

•En puntuaciones diferenciales:

$$y' = r_{xy}\frac{S_y}{S_x}x_i$$

•En puntuaciones típicas:

$$Z'_y = r_{xy}Z_x$$

-Estimación de la puntuación verdadera en el criterio: intervalo de confianza:

•Error típico de estimación:

$\sigma_{\text{est.}} = S_{yx} = S_y\sqrt{1 - r_{xy}^2}$ para puntuaciones directas y diferenciales.

$\sigma_{\text{est.}} = S_{yx} = \sqrt{1 - r_{xy}^2}$ para puntuaciones típicas.

•Intervalo de confianza:

$$Y = Y' \pm \sigma_{\text{est.}} Z_{nc} = Y' \pm \sqrt{1 - r_{xy}^2} \cdot Z_{nc}$$

$$y = y' \pm \sigma_{\text{est.}} Z_{nc} = y' \pm S_y \sqrt{1 - r_{xy}^2} \cdot Z_{nc}$$

$$Z_y = Z'_y \pm \sigma_{\text{est.}} Z_{nc} = Z'_y \pm S_y \sqrt{1 - r_{xy}^2} \cdot Z_{nc}$$

C) Relación de la validez con otras variables:

•Validez y fiabilidad:

$$\frac{r_{xy}}{R_{xy}} = \sqrt{\frac{r_{xx} r_{yy}}{R_{xx} R_{yy}}}$$

r=xy coeficiente de validez inicial.

R=xy nuevo coeficiente de validez.

r=' coeficiente de fiabilidad inicial del test.

R=xx nuevo coeficiente de fiabilidad del test.

r=yy coeficiente de fiabilidad inicial del criterio.

= nuevo coeficiente de fiabilidad del criterio.

•Validez y longitud:

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}}; \qquad n = \frac{1 - r_{xx}}{\frac{r_{xy}^2}{R_{xy}^2} - r_{xx}}$$

•Validez y variabilidad:

$$R_{xy} = \frac{S_x r_{xy}}{\sqrt{S_x^2 r_{xy}^2 + (1 - r_{xy}^2) S_x^2}}$$

Sx = varianza del grupo con variabilidad mayor o menor.

Sx = varianza en el test original.

10.1.3.2. Ejercicios resueltos

1. A partir de los resultados obtenidos en la prueba del apartado 10.1.2., y teniendo en cuenta que las puntuaciones de la columna Y del cuadro de datos (cuadro 10.1) corresponden a la medida de un criterio externo:

   a) Calcular el coeficiente de validez de la prueba.

   b) Suponiendo que el instrumento de medida se ha elaborado con fines predictivos y que la medida del criterio externo se hubiera realizado tras un período de un año después de la aplicación de la citada prueba:

      1. ¿Qué puntuación directa se le predeciría a un sujeto en el criterio, sabiendo que en la prueba obtuvo 20 puntos?

      2. ¿Entre qué límites se encontrará la puntuación verdadera del mencionado sujeto a un nivel de confianza del 98%?

   c) ¿Qué valor alcanzaría el coeficiente de validez de la prueba si su coeficiente de fiabilidad se incrementara de 0,93 (valor obtenido en el apartado anterior) a 0,95?

   d) Calcular la nueva validez de la prueba si se duplicara su longitud.

   e) ¿Qué validez alcanzará la prueba si al aplicarla a una muestra de 100 sujetos hemos obtenido una S, = 9, sabiendo que en la muestra de nuestro estudio, formada por 46 sujetos, la desviación típica fue de 10,74 (valor obtenido en los ejercicios del apartado anterior)?

SOLUCIÓN:

a) Para obtener el coeficiente de validez debemos calcular un coeficiente de correlación entre las puntuaciones totales en el test y en el criterio. El coeficiente de correlación dependerá de las características de las variables medidas. En nuestro caso, suponemos que las variables son continuas y distribuidas normalmente, por lo que utilizaremos el coeficiente de correlación de Pearson.

   La fórmula es la siguiente:

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

   En el siguiente cuadro se ofrecen los resultados de los datos correspondientes para sustituir en la fórmula anterior.

CUADRO 10.3. Resultados para el cálculo de la validez

Sujetos	X_i	Y_i	X_i^2	Y_i^2	*XY*
1	21	4	441	16	84
2	36	8	1.296	64	288
3	16	2	256	4	32
4	34	8	1.156	64	272
5	23	4	529	16	92
6	25	5	625	25	125
7	7	2	49	4	14
8	5	1	25	1	5
9	33	6	1.089	36	198
10	22	5	484	25	110
11	23	4	529	16	92
12	20	3	400	9	60
13	28	4	784	16	112
14	9	5	81	25	45
15	41	9	1.681	81	369
16	12	3	144	9	36
17	24	6	576	36	144

Sujetos	X_i	Y_i	X_i^2	Y_i^2	*XY*
18	7	1	49	1	7
19	29	6	841	36	174
20	31	8	961	64	248
21	28	7	784	49	196
22	7	2	49	4	14
23	37	8	1.369	64	296
24	21	3	441	9	63
25	41	9	1.681	81	369
26	28	5	784	25	140
27	31	4	961	16	124
28	40	8	1.600	64	320
29	36	6	1.296	36	216
30	23	6	529	36	148
31	14	4	196	16	56
32	20	3	400	9	60
33	19	4	361	16	76
34	15	4	225	16	60
35	39	9	1.521	81	351
36	8	3	64	9	24
37	25	5	625	25	125
38	29	7	841	49	203
39	35	7	1.225	49	245
40	24	5	576	25	120
41	44	10	1.936	100	440
42	8	5	64	25	40
43	8	3	64	9	24
44	16	4	256	16	64
45	12	4	144	16	48
46	25	6	625	36	150
Σ	1.097	235	30.613	1.429	6.479

$$\sum X = 1.079$$
$$\sum Y = 235$$
$$\sum X^2 = 30.613$$
$$\sum Y^2 = 1.429$$
$$\sum XY = 6.479$$

$$r_{xy} = \frac{46 \cdot 6.479 - (1.079)(235)}{\sqrt{(46 \cdot 30.613 - 1.079^2)(46 \cdot 1.429 - 235^2)}} =$$
$$= \frac{297.988 - 253.565}{\sqrt{(1.408.198 - 1.164.241)(65.734 - 55.225)}} =$$
$$= \frac{44.423}{\sqrt{243.957 \times 10.509}} = 0{,}877$$

El resultado obtenido (xx = 0,877) nos muestra la alta validez que tiene la prueba con el criterio considerado.

b)1. Para la resolución de esta cuestión debemos utilizar la ecuación de regresión para predecir la puntuación directa del sujeto en el criterio:

$$Y' = r_{xy} \frac{S_y}{S_x} X_i - r_{xy} \frac{S_y}{S_x} \bar{X} + \bar{Y} = r_{xy} \frac{S_y}{S_x} (X_i - \bar{X}) + \bar{Y}$$

De estos valores tenemos los siguientes:

X= 20.

rxy = 0,877 (obtenido en el apartado a) de este ejercicio)

S'= 10,74 (obtenido en los ejercicios del apartado anterior)

$$\bar{X} = \frac{\sum X_i}{N} = \frac{1.079}{46} = 23,45$$

$$\bar{Y} = \frac{\sum Y_i}{N} = \frac{235}{46} = 5,10$$

$$S_y = \sqrt{\frac{\sum Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2} = \sqrt{\frac{1.429}{46} - \left(\frac{235}{46}\right)^2} = 2,23$$

Sustituimos en la primera fórmula:

$$Y' = 0,877\frac{2,23}{10,74}(20 - 23,45) + 5,1 = 4,48$$ puntuación que se le predice al sujeto en el criterio.

En el caso de que se nos hubiera pedido la estimación en puntuaciones diferenciales o típicas, pasaríamos, en primer lugar, la puntuación directa en la prueba a puntuaciones diferenciales o típicas, y aplicaríamos después la ecuación de regresión correspondiente.

2. A partir de la puntuación predicha que acabamos de calcular (Y = 4,48), tenemos que obtener los límites entre los que podemos asegurar se encontrará la puntuación verdadera del sujeto a un nivel de confianza del 98%. Para ello utilizaremos la fórmula de cálculo del intervalo de confianza para puntuaciones directas:

$$Y = Y' \pm \sigma_{\text{med.}} \cdot Z_{nc}$$

$$\sigma_{\text{est.}} = S_y\sqrt{1 - r_{xy}^2} = 2,23\sqrt{1 - 0,877^2} = 1,07$$

Z = 98% = 2,33 (véanse tablas de curva normal)

$$Y = 4,48 \pm 1,07 \cdot 2,33 = \left[\begin{array}{l} 1,99 \\ 6,97 \end{array}\right.$$

Entre 1,99 y 6,97 se encontrará la verdadera puntuación del sujeto a un nivel de confianza de 98%.

c) Aplicamos la fórmula que relaciona validez y fiabilidad:

$$\frac{r_{xy}}{R_{xy}} = \sqrt{\frac{r_{xx}r_{yy}}{R_{xx}R_{yy}}}$$

De estos datos conocemos los siguientes:

r_{xy} = 0,877.
r_{xx} = 0,93.
R_{xx} = 0,95.
R_{xy} = valor que vamos a obtener.

r=YY fiabilidad del criterio. Aunque este valor se desconoce, en este caso el problema pue- de resolverse, ya que no se nos dice que varíe este coeficiente de fiabilidad. Por tanto, se entiende que rn,= Así pues, desaparecen del numerador y denominador.

Sustituyendo:

$$\frac{0,877}{R_{xy}} = \sqrt{\frac{0,93 \cdot r_{yy}}{0,95 \cdot R_{yy}}}; \ \frac{0,877}{R_{xy}} = 0,989$$

$$R_{xy} = \frac{0,877}{0,989} = 0,886 \text{ valor del nuevo coeficiente de validez.}$$

Como vemos por el resultado, el incremento de la fiabilidad ha motivado el que la validez sea también más alta, pasando de 0,877 a 0,886.

d)Aplicamos la fórmula que relaciona validez y longitud, teniendo en cuenta que n = 2, ya que la prueba duplica su longitud:

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}} = \frac{0,877}{\sqrt{\frac{1 - 0,93}{2} + 0,93}} = 0,893$$

La validez se incrementa, pasando de 0,877 a 0,893.

e)Aplicamos la fórmula que relaciona validez y variabilidad:

$$R_{xy} = \frac{S_x r_{xy}}{\sqrt{s_x^2 r_{xy}^2 + (1 - r_{xy}^2) S_x^2}}$$

$S_x = 9.$
$r_{xy} = 0{,}877.$
$s_x = 10{,}74.$

Sustituyendo:

$$R_{xy} = \frac{9 \cdot 0{,}877}{\sqrt{9^2 \cdot 0{,}877^2 + (1 - 0{,}877^2)10{,}74^2}} = \frac{7{,}893}{\sqrt{62{,}289 + 26{,}64}} =$$

$$= \frac{7{,}893}{9{,}43} = 0{,}837 \text{ valor del nuevo coeficiente de validez.}$$

En este caso, el coeficiente de validez se ha reducido de 0,877 a 0,837, al calcularse en una muestra de menor variabilidad.

10.1.4. Análisis de ítems

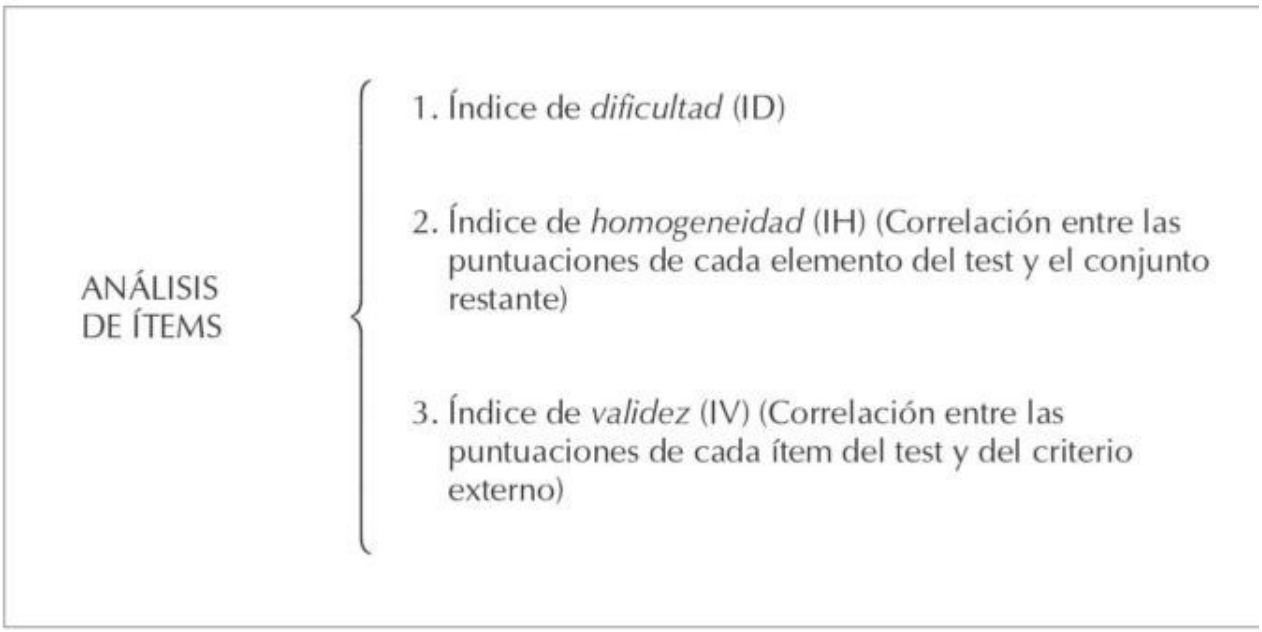


Figura 10.3. Análisis de ítems y sus indicadores.

10.1.4.1. Formulario

A) Índice de dificultad (ID):

- $ID = \frac{A}{n}$

A = número de aciertos.

n = número de intentos.

•Procedimiento abreviado:

$$ID = \frac{A_s + A_i}{n_s + n_i}$$

los subíndices seiindican grupos superior e inferior, respectivamente.

•En caso de elementos de elección múltiple:

$$ID = \frac{A - \frac{E}{a-1}}{n}$$

E = errores.

a = número de alternativas u opciones.

B) Índice de homogeneidad (IH):

IH= correlación entre el ítem y el resto del test (rt,n,, resto del test)

IH=(1),test(T), corregido con la fórmula:

$$IH = \frac{r_{IT} s_T - s_I}{\sqrt{s_T^2 + s_I^2 - 2 r_{IT} s_T s_I}}$$

NOTA: Esta corrección sólo debe realizarse cuando el número de elementos de la prueba es pequeño.

C) Indice de validez (IV):

$$IV = r_{\text{ítem, criterio}}$$

10.1.4.2. Ejercicios resueltos

1. A partir de los resultados obtenidos en la prueba, realice el análisis de los ítems o elementos números 15, 20Y 30, calculando sus correspondientes índices y valorando los resultados para una adecuada selección de los elementos.

SOLUCIÓN:

El análisis de ítems debe realizarse de todos y cada uno de los elementos de la prueba a fin de seleccionar los más adecuados y conseguir una prueba que reúna las mejores características técnicas. No obstante, para la ejemplificación hemos seleccionado únicamente tres elementos de la prueba, ya que el procedimiento sería igual para todos.

A continuación ofrecemos el cuadro de puntuaciones obtenidas por cada uno de los 46 sujetos en los tres ítems, en el total de la prueba y en el criterio. Nos fijaremos en este cuadro 10.4 para todos los cálculos que se van a realizar en este apartado.

CUADRO 10.4. Resultados en los ítems 15, 20 y 30, puntuaciones totales (X) y en el criterio (Y)

Sujetos	*Ítem 15*	*Ítem 20*	*Ítem 30*	X_t	*Y*
1	+	+		21	4
2	+	+	+	36	8
3		+		16	2
4	+	+		14	8
5	+	+		23	4
6	+	+		25	5
7		+		7	2
8		+		5	1
9		+	+	33	4
10		+		22	5
11		+		23	4
12		+		20	3
13	+		+	28	4
14		+		9	5
15	+	+	+	41	9
16		+		12	3
17		+		24	6
18		+		7	1
19		+		29	6
20	+	+		31	8
21	+	+		28	7
22				7	2
23	+	+	+	37	8
24				21	3
25	+	+	+	41	9
26	+			28	5
27		+	+	31	4
28	+	+	+	40	8
29	+		+	36	6
30		+	+	23	6
31		+		14	4

Sujetos	*Ítem 15*	*Ítem 20*	*Ítem 30*	X_t	*Y*
32		+		20	3
33				19	4
34		+		15	4
35	+	+	+	39	9
36		+		8	3
37		+		25	5
38		+		29	7
39	+	+	+	35	7
40		+		24	5
41	+	+		44	10
42		+		8	5
43		+		8	3
44	+	+		16	4
45		+		12	4
46	+	+		25	6
Σ	19	39	12	1.079	235

•Comenzaremos calculando el índice de dificultad aplicando la fórmula:

$$ID = \frac{A}{n}$$

A = número de aciertos.

n = número de intentos.

Ítem 15: $ID = \frac{19}{46} = 0{,}41$ es un ítem difícil, pero está próximo al límite que se considera normal (0,44).

Ítem 20: $ID = \frac{39}{46} = 0{,}874$ es un ítem muy fácil.

Ítem 30: $ID = \frac{12}{46} = 0,26$ es un ítem bastante difícil, de mayor dificultad que el ítem 15.

•Índice de homogeneidad:

Para calcular este índice utilizaremos una correlación entre las puntuaciones de cada sujeto en el ítem y en el resto de la prueba o en el total de la misma, realizando, en este caso, la correlación correspondiente, sobre todo cuando el número de elementos de la prueba es pequeño.

El tipo de correlación a utilizar dependerá de las características de los datos de la prueba y del ítem. En nuestro caso, utilizaremos la correlación biserial puntual, ya que una de las variables es continua y la otra dicotómica (ítems). No procede utilizar procedimientos abreviados, ya que el número de sujetos es pequeño.

Así pues, procederemos a aplicar el coeficiente de correlación biserial puntual entre el ítem y el total de la prueba. Realizaremos posteriormente la corrección, aunque, en este caso, no creemos que altere mucho su valor, ya que el número de elementos de la prueba (50) no es muy pequeño.

ÍTEM 15:

$$r_{bp} = \frac{\overline{x}_p - \overline{x}_t}{s_t}\sqrt{p/q}$$

Xp = media de los sujetos que aciertan el ítem.

Xr = media total.

Sr = desviación típica de la prueba.

p = proporción de aciertos.

q = 1 - p.

$$\overline{X}_t = \frac{\sum X_i}{N} = \frac{1.079}{46} = 23,456$$

Sr = 10,74 (ya calculada anteriormente)

$$p = \frac{19}{46} = 0,41$$

$$q = 1 - 0,41 = 0,59$$

La Xp la calcularemos sumando las puntuaciones totales en la prueba de aquellos sujetos que acertaron el ítem y dividiendo entre el número de éstos. Para el cálculo de este valor se suelen utilizar procedimientos abreviados agrupando las puntuaciones totales en el test en intervalos y viendo las frecuencias de aciertos y de errores, etc. No obstante, teniendo en cuenta que el número de sujetos no es muy elevado, lo realizaremos directamente:

$$X_p = \frac{\sum X_p}{n_p} = \frac{588}{19} = 30,947$$

Sustituyendo los valores:

$$r_{bp} = \frac{30,947 - 23,456}{10,74}\sqrt{0,41 / 0,59} = \frac{7,491}{10,74} \cdot 0,83 = 0,578$$

Aplicamos la fórmula de corrección:

$$IH = \frac{r_{IT} S_T - S_I}{\sqrt{S_T^2 + S_I^2 - 2 r_{IT} S_T S_I}}$$

T = prueba en su conjunto.
I = elemento de la prueba.
$r_{IT} = 10,74$ (ya calculado anteriormente)
$S_I = \sqrt{pq}$ (por ser elementos dicotómicos)
$S_I = \sqrt{0,41 \cdot 0,59} = 0,49.$
$S_I^2 = 0,24.$

Sustituyendo:

$$IH = \frac{0,578 \cdot 10,74 - 0,49}{\sqrt{10,74^2 + 0,49^2 - 2 \cdot 0,578 \cdot 10,74 \cdot 0,49}} = \frac{5,71}{10,46} = 0,545$$

El índice de homogeneidad de este ítem es bastante elevado, muy superior al mínimo exigido, por lo que, según este criterio, debería ser seleccionado para la

prueba.

ÍTEM 20:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{p/q}$$

$$\left.\begin{array}{l}\bar{X}_t = 23{,}456 \\ S_t = 10{,}74\end{array}\right] \text{Ya obtenidos anteriormente.}$$

$$p = \frac{39}{46} = 0{,}847$$

$$q = 1 - 0{,}847 = 0{,}163$$

$$\bar{X}_p = \frac{\sum X_p}{n_p} = \frac{920}{39} = 23{,}59$$

Sustituyendo:

$$r_{bp} = \frac{23{,}59 - 23{,}456}{10{,}74}\sqrt{0{,}847 / 0{,}163} = \frac{0{,}134}{10{,}74} \cdot 2{,}279 = 0{,}028$$

Corrección:

$$IH = \frac{r_{IT}S_T - S_I}{\sqrt{S_T^2 + S_I^2 - 2r_{IT}S_TS_I}}$$

$$r_{IT} = 0{,}028$$

$$S_T = 10{,}74$$

$$S_I = \sqrt{pq} = \sqrt{0{,}847 \cdot 0{,}163} = 0{,}37$$

$$IH = \frac{0{,}028 \cdot 10{,}74 - 0{,}37}{\sqrt{10{,}74^2 + 0{,}37^2 - 2 \cdot 0{,}028 \cdot 10{,}74 \cdot 0{,}37}} = \frac{-0{,}07}{10{,}73} = -0{,}006$$

Este índice de homogeneidad es muy bajo y negativo, por lo que este ítem debería ser eliminado.

ÍTEM 30:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_t}{S_t}\sqrt{p/q}$$

$$\left.\begin{array}{l}\bar{X}_t = 23,456 \\ S_t = 10,74\end{array}\right] \text{Ya obtenidos.}$$

$$p = \frac{12}{46} = 0,26$$

$$q = 1 - 0,26 = 0,74$$

$$\bar{X}_p = \frac{\sum X_p}{n_p} = \frac{450}{12} = 37,5$$

Sustituyendo:

$$r_{bp} = \frac{37,5 - 23,456}{10,74}\sqrt{0,26 / 0,74} = \frac{14,044}{10,74} \cdot 0,59 = 0,77$$

Corrección:

$$IH = \frac{r_{IT}S_T - S_I}{\sqrt{S_T^2 + S_I^2 - 2r_{IT}S_TS_I}}$$

$$r_{IT} = 0,77$$

$$S_T = 10,74$$

$$S_I = \sqrt{pq} = \sqrt{0,26 \cdot 0,74} = 0,438$$

$$IH = \frac{0,77 \cdot 10,74 - 0,438}{\sqrt{10,74^2 + 0,438^2 - 2 \cdot 0,77 \cdot 10,74 \cdot 0,438}} = \frac{7,83}{10,40} = 0,752$$

El ítem 30 tiene un índice de homogeneidad muy alto. Debería ser incluido en la prueba según este criterio.

•Índice de validez:

La obtención del índice de validez se realiza mediante una correlación entre las puntuaciones de cada uno de los elementos y el criterio (véase cuadro 10.4). El

procedimiento es prácticamente igual que para el cálculo del índice de homogeneidad, cambiando las puntuaciones totales de la prueba por las del criterio (columna Í). En nuestro caso, deberemos utilizar igualmente el coeficiente de correlación biserial puntual, ya que el criterio es una variable continua y el ítem es dicotómico.

ÍTEM 15:

$$r_{bp} = \frac{\bar{Y}_p - \bar{Y}}{S_y} \cdot \sqrt{p / q}$$

Y = media de las puntuaciones del criterio. En este caso, Xpasa a ser Y, que es como hemos denominado a la variable.

$$\left.\begin{array}{l} \bar{Y} = \dfrac{\sum Y}{N} = \dfrac{235}{46} = 5{,}1 \\ S_Y = 2{,}23 \end{array}\right] \text{Ya obtenido anteriormente.}$$

$$\left.\begin{array}{l} p = 0{,}41 \\ q = 0{,}59 \end{array}\right] \text{Ya obtenidos en el cálculo del índice de homogeneidad de este ítem.}$$

$$\bar{Y}_p = \frac{\sum Y_p}{n_p} = \frac{129}{19} = 6{,}79$$

Y = puntuación en el criterio de los sujetos que aciertan el ítem.

Sustituyendo:

$$r_{bp} = \frac{6{,}79 - 5{,}1}{2{,}23} \cdot \sqrt{0{,}41 / 0{,}59} = \frac{1{,}69}{2{,}23} \cdot 0{,}83 = 0{,}628$$

El índice de validez del ítem 15 es bastante alto y, por tanto, adecuado para la prueba.

ÍTEM 20:

$$r_{bp} = \frac{\bar{Y}_p - \bar{Y}}{S_y} \cdot \sqrt{p / q}$$

$$\left.\begin{array}{l}\bar{Y} = 5,1 \\ S_Y = 2,23\end{array}\right] \text{Ya calculados.}$$

$$p = \frac{39}{46} = 0,847$$

$$q = 0,163$$

$$\bar{Y}_p = \frac{\sum Y_p}{n_p} = \frac{211}{39} = 5,41$$

Sustituyendo:

$$r_{bp} = \frac{5,41 - 5,1}{2,23} \cdot \sqrt{0,847 / 0,163} = \frac{0,31}{2,23} \cdot 2,279 = 0,319$$

El índice de validez del ítem 20 es adecuado, superior al límite de 0,20 recomendado.

ÍTEM 30:

$$r_{bp} = \frac{\bar{Y}_p - \bar{Y}}{S_y} \cdot \sqrt{p / q}$$

$$\left.\begin{array}{l}\bar{Y} = 5,1 \\ S_Y = 2,23\end{array}\right] \text{Ya calculados.}$$

$$p = \frac{12}{46} = 0,26$$

$$q = 0,74$$

$$\bar{Y}_p = \frac{\sum Y_p}{n_p} = \frac{84}{12} = 7$$

Sustituyendo:

$$r_{bp} = \frac{7 - 5,1}{2,23} \cdot \sqrt{0,26 / 0,74} = \frac{1,9}{2,23} \cdot 0,59 = 0,50$$

El ítem 30 tiene un buen índice de validez.

•Valoración de los resultados obtenidos en los ítems analizados:

	ÍTEMS		
	15	20	30
ID	0,41	0,847	0,26
IH	0,545	–0,006	0,752
IV	0,628	0,319	0,50

Teniendo en cuenta los resultados, podemos observar cómo los tres ítems tienen índices de validez aceptables (superiores a 0,20) e incluso muy aceptables. De ello podríamos deducir que los tres ítems deberían ser incluidos en la prueba. No obstante, si analizamos el índice de homogeneidad, nos encontramos con que el ítem 20 tiene un índice negativo, por lo que debería ser eliminado de la prueba. Los índices de dificultad son menos decisivos a la hora de seleccionar un elemento concreto, ya que una prueba debe tener elementos de distintos niveles de dificultad en porcentajes diferentes. De aquí que, en nuestro caso, nos encontremos con tres ítems de distinto nivel de dificultad, desde muy fácil a muy difícil.

Así pues, en resumen, incluiríamos los ítems 15 y 30 y eliminaríamos el ítem 20 por su índice de homogeneidad negativo.

OBSERVACIONES:

1. En la obtención de los índices de homogeneidad y de validez, cuando tenemos muchos ítems y un número representativo de sujetos, se suele utilizar, en muchos casos, el coeficiente de correlación tetracórico, dicotomizando una o las dos variables (generalmente por la mediana) para evitar el elevado número de cálculos que es necesario realizar con otros coeficientes de correlación.

   El coeficiente de corrrelación tetracórico, además de la facilidad para calcularse utilizando la fórmula, permite obtener su valor a través de tablas que simplifican mucho más el proceso. A pesar de todo, este problema queda en un segundo plano con el uso de los programas de ordenador correspondientes.

2. Como es lógico, los índices de homogeneidad y validez nos permiten conocer la proporción de varianza que cada uno de los ítems explican de la prueba o del criterio, teniendo en cuenta que son coeficientes de correlación. Por ello, muchas de las cuestiones o ejercicios que se pueden plantear no hacen referencia directa a estos índices, sino a la cantidad de varianza que explican.

zo.1.5. Tipificación

1. *Cronológicas o de edad.*	1.1. Cociente intelectual.	
	1.2. Cociente de instrucción.	
	1.3. Cociente de rendimiento intelectual.	
2. *Cuantiles.*	2.1. Centiles	
	2.2. Deciles	
	2.3. Cuartiles	
3. *Típicas.*	3.1. Puntuaciones típicas (Z)	
	3.2. Puntuaciones típicas normalizadas (Z_n)	
	3.3. Puntuaciones típicas derivadas	– CEEB
		– *S*
		– *T*
		– Eneatipos
		– Pentas

Figura 10.4. Tipificación: Escalas individuales.

10.1.5.1. Formulario

A) Puntuaciones cronológicas:

1.Cociente intelectual:

$$CI = \frac{EM}{EC} \cdot 100$$

2.Cociente de instrucción:

$$C.\ Instr. = \frac{E.\ Instr.}{EC} \cdot 100$$

3.Cociente de rendimiento intelectual:

$$C.R.int. = \frac{E.Instr.}{EM} \cdot 100$$

Donde:

EM= edad mental.

EC= edad cronológica.

E.Instr. = edad de instrucción.

B) Cuantiles (centiles, deciles y cuartiles):

Centil: $$C_m = L_{\text{inf.}} + \frac{m \cdot N/100 - f_{a(I-1)}}{f_i} \cdot C$$

Decil: $$D_m = L_{\text{inf}} + \frac{m \cdot N/10 - f_{a(I-1)}}{f_i} \cdot C$$

Cuartil: $$Q_m = L_{\text{inf.}} + \frac{m \cdot N/4 - f_{a(I-1)}}{f_i} \cdot C$$

Donde:

m: cuantil que deseamos obtener.

f.: límite inferior del intervalo donde se encuentra el cuantil (decil, centil o cuartil) N. número total de casos.

f(11):frecuencia acumulada del intervalo anterior a aquel en que se encuentra el cuantil. c amplitud del intervalo.

C) Puntuaciones típicas y derivadas:

Típicas: $$Z_i = \frac{X_i - \bar{X}}{S}$$

Derivadas:

•T= 50 + (10 • Z)

•S=50+(20•Z)

•CEEB=500+(100.2)

•Penta:P=3+(1•Z~

•Eneatipos: E= 5 + (2 • Z)

Baremos en pentas:

Los cinco pentas se hallan delimitados por las puntuaciones Z.•-1,5; - 0,5; 0,5; 1,5.

Baremos de eneatipos:

Los nueve eneatipos se hallan delimitados por las puntuaciones Z.•-1,75; - 1,25; - 0,75;

- 0,25; 0,75; 1,25; 1,75.

10.1.5.2. Ejercicios resueltos

1.Elaborar los baremos de la prueba del estudio en centiles, puntuaciones típicas y típicas normalizadas, eneatipos y pentas.

SOLUCIÓN:

De los tipos de puntuaciones tipificadas que hemos propuesto en el esquema y formulario, son los cuantiles y las puntuaciones típicas y derivadas las de uso más frecuente, mientras que no se suelen encontrar en la actualidad baremos en puntuaciones cronológicas. Por esto, en el ejemplo, nos referimos únicamente a las ya mencionadas.

En primer lugar, calcularemos los cuantiles. Dado que los deciles y cuartiles se calculan de forma similar a los centiles, ya que en realidad vienen a constituir parte de éstos, calcularemos solamente los centiles. A título de ejemplo, presentaremos los cálculos detallados de dos de ellos, si bien se presentan los resultados de los seleccionados por su mayor uso (centil, 1, 5, 10, 20,..., 99).

Para el cálculo, agruparemos las puntuaciones totales en la prueba (Z) en intervalos de amplitud 5. De esta forma obtenemos ocho intervalos: (44 - 5 + 1)/5 = 8. Se presentan igualmente las frecuencias absolutas (f) y acumuladas (f) correspondientes a cada intervalo. A partir de estos datos calcularemos los centiles de nuestro interés,

detallando el cálculo del centil 10 y del 90. Como sabemos, en cualquier centil pretendemos encontrar la puntuación directa que deja debajo de sí el correspondiente porcentaje de casos.

I	f_i	f_a
5-9	8	8
10-14	3	11
15-19	4	15
20-24	10	25
25-29	8	33
30-34	4	37
35-39	5	42
40-44	4	46
	$N = 46$	

GENTIL 10:

Calculamos el lugar que ocupa en la serie el sujeto o puntuación que deja debajo de sí el 10% de los casos:

46 sujetos ——— 100% de los casos
x sujetos ——— 10% de los casos

$$x(\text{lugar que ocupa en la serie}) = \frac{46 \times 10}{100} = 4{,}6$$

Así, la puntuación correspondiente al centil 10 ocupará el lugar 4,6 de la serie, empezando por la puntuación inferior. Mirando en la columna de f vemos que este sujeto se encuentra en el intervalo 5-9, puesto que hay 8 puntuaciones de sujetos, que van desde la puntuación más baja de la distribución, correspondiente al primer sujeto, hasta el sujeto octavo.

A continuación aplicamos la fórmula del centil:

$$C_m = L_{\text{inf.}} + \frac{m \cdot N/100 - f_{a(I-1)}}{f_i} \cdot C$$

$$C_{10} = 4{,}5 + \frac{10 \cdot 46/100 - 0}{8} \cdot 5 = 4{,}5 + 2{,}875 = 7{,}375$$

GENTIL 90:

46 sujetos ——— 100% de los casos
x sujetos ——— 90% de los casos

$$x = \frac{46 \times 90}{100} = 41{,}4 \text{ lugar que ocupa la puntuación correspondiente al } C_{90}.$$

Columna de f -* el lugar que ocupa el sujeto 41,4 se encuentra en el intervalo 3539, donde se encuentran los 42 primeros sujetos y, en consecuencia, el 41,4:

$$C_m = L_{\text{inf.}} + \frac{m \cdot N/100 - f_{a(I-1)}}{f_i} \cdot C$$

$$C_{90} = 34{,}5 + \frac{90 \cdot 46/100 - 37}{5} \cdot 5 = 34{,}5 + 4{,}4 = 38{,}9$$

D_m			1	2	3	4	5	6	7	8	9	
C_m	1	5	10	20	30	40	50	60	70	80	90	99
X_i	4,78	5,93	7,37	11,5	18	21	23,5	26,12	29	34,25	38,9	43,92
X_i redondeada	5	6	7	11	18	21	23	26	29	34	39	44

En la tabla se presentan los centiles y sus puntuaciones directas correspondientes. Hemos redondeado estas puntuaciones, ya que la prueba no permite obtener valores decimales.

Teniendo en cuenta que los centiles 10, 20, 30, etc., equivalen a los deciles 1, 2, 3, etc., hemos incluido también en la primera fila estos cuantiles.

PUNTUACIONES TÍPICAS Y TÍPICAS NORMALIZADAS:

Para la obtención de estas puntuaciones partimos de la misma tabla de intervalos y frecuencias que hemos utilizado para los centiles y que reproducimos a continuación.

En este caso obtendremos las puntuaciones (Zy ZZ que corresponden a los límites superiores de los intervalos.

I	f	f_a	P_a	Z_n	Z
5-9	8	8	17,39	–0,94	–1,298
10-14	3	11	23,91	–0,71	–0,833
15-19	4	15	32,60	–0,45	–0,367
20-24	10	25	54,34	0,11	0,097
25-29	8	33	71,73	0,57	0,563
30-34	4	37	80,43	0,86	1,028
35-39	5	42	91,30	1,36	1,494
40-44	4	46	100	–	1,959

•Las puntuaciones Z se hallan a partir de los porcentajes acumulados (P). Esta columna se ha obtenido calculando el porcentaje de los casos que representa cada una de las frecuencias acumuladas respecto del total (N= 46). Así, el primer porcentaje será igual a 8 x 100/46 = 17,39, y así continuaremos hasta llegar ala última frecuencia acumulada que representa el 100% de los casos.

Con los datos de esta columna (P) vamos a las tablas de curva normal, viendo la puntuación típica que deja debajo de sí ese porcentaje de casos. Así, en el primer intervalo, la puntuación Z que corresponde al límite superior del intervalo (X = 9,5), según las tablas de curva normal, será - 0,94. Deja debajo de sí el 17,94% de los casos. La puntuación es negativa al estar a la izquierda de la distribución. Evidentemente, para mirar las tablas iremos a la columna de área menor con valor 0,1794 = (17,94/100). Así sucesivamente obtendremos las Z que corresponden a los límites superiores de los intervalos. Si queremos obtener la puntuación típica normalizada que corresponde a cualquier otra puntuación directa, únicamente tendremos que ver el porcentaje de casos que deja debajo de sí esa puntuación, es decir, calcular el centil que le corresponde, y después ir a las tablas de curva normal con ese porcentaje, buscando la puntuación típica correspondiente.

•Las puntuaciones típicas (4 se obtienen directamente a través de la fórmula:

$$Z_i = \frac{X_i - \bar{X}}{S}$$

En la columna Zde la tabla se presentan las puntuaciones típicas correspondientes a los límites superiores de los intervalos. Así, teniendo en cuenta que la media (X) la de la distribución es igual a 23,45 y la desviación típica igual a 10,74 (valores calculados anteriormente), la Z. correspondiente al límite superior del primer intervalo (9,5) sería:

$$Z_i = \frac{9,5 - 23,45}{10,74} = -1,298$$

Si queremos obtener la puntuación típica correspondiente a cualquier otra puntuación, independientemente de los límites superiores de los intervalos, no tendremos más que aplicar la mencionada fórmula.

Las diferencias entre las puntuaciones Z y Zque podemos observar en la tabla se deben a la ligera discrepancia entre la distribución empírica y la distribución normal, si bien podemos afirmar que ambas distribuciones no difieren significativamente (comprobación realizada a través de la prueba de xz). No obstante, cuanto mayor sea el ajuste, menor será la diferencia que habrá entre las puntuaciones típicas y típicas normalizadas. Generalmente, con mayor número de sujetos de la muestra se dará un mayor ajuste.

Como las puntuaciones Zy Z plantean ciertos inconvenientes para su interpretación, por la existencia de valores negativos y decimales, se suelen utilizar las puntuaciones típicas derivadas, cuyas escalas más conocidas son las S (X = 50 y S= 20), T(X= 50 y S= 10) y CEEB (,Y= 500 y S= 100).

Para la transformación de cualquier puntuación típica en una de estas escalas, únicamente habrá que multiplicar dicha puntuación por la desviación típica de la escala correspondiente y sumarle a este producto la media de la escala. Así por ejemplo, para transformar el primer valor de Zde la tabla de datos en cada una de estas tres escalas, procederíamos como sigue:

$$S = 50 + 20\ (-1,298) = 24,04$$
$$T = 50 + 10\ (-1,298) = 37,02$$
$$CEEB = 500 + 100\ (-1,298) = 370,2$$

ENEATIPOS Y PENTAS

Para la obtención de las puntuaciones directas que limitan los 9 o 5 rangos de estas escalas, tenemos que partir de las puntuaciones típicas que los limitan y que, como sabemos, son valores fijos.

Así, la escala de eneatipos divide la escala en 9 rangos cuyos límites en Zson los que podemos observar en la primera fila de Z.-

Z		–1,75		–1,25		–0,75		–0,25		0,25		0,75		1,25		1,75	
Eneatipo	1º		2º		3º		4º		5º		6º		7º		8º		9º
X_i		4,655		10,025		15,395		20,765		26,135		31,5		36,875		42,245	

Igualmente podemos observar los límites de los pentas:

Z	–1,5	–0,5	0,5	1,5	
Pentas	1º	2º	3º	4º	5º
X_i	7,34	18,08	28,82	39,56	

Para la obtención de las correspondientes puntuaciones directas, únicamente aplicaremos la fórmula:

$$Z_i = \frac{X_i - \bar{X}}{S}$$

donde despejaremos el valor X.

Así, para el primer eneatipo:

$Z_i = -1{,}75$
$\bar{X} = 23{,}45$
$S = 10{,}74$

$$-1{,}75 = \frac{X_i - 23{,}45}{10{,}74}$$

$$X_i = 23{,}45 + (10{,}74)(-1{,}75) = 4{,}655$$

Primer penta:

$$-1{,}5 = \frac{X_i - 23{,}45}{10{,}74}$$

$$X_i = 23{,}45 + (10{,}74)(-1{,}5) = 7{,}34$$

La obtención de las puntuaciones directas correspondientes a los límites de cada uno de los rangos de las dos escalas se realizaría por el mismo procedimiento, sustituyendo el valor de Z.En las tablas anteriores se presentan los resultados obtenidos de los cálculos (fila de X.)

También se pueden calcular las puntuaciones directas correspondientes a las puntuaciones típicas límites de los rangos, como si la distribución siguiera exactamente la curva normal. Para ello, deberíamos ver en la tabla de curva normal el porcentaje de casos que deja debajo de sí cada una de las Z que limitan los eneatipos o pentas y, a continuación, calcular, a partir de la distribución de datos, la X que deja debajo de sí tal

porcentaje. En este caso, los valores obtenidos no deben discrepar sustancialmente de los obtenidos a partir de la distribución empírica, ya que ambas distribuciones no difieren significativamente. No obstante, como ya hemos señalado anteriormente, a mayor ajuste menor discrepancia, y es de esperar que esto se produzca con un mayor número de sujetos, tal como se suele realizar a la hora de obtener los baremos.

Además de la elaboración de baremos, se pueden plantear cuestiones referidas a la situación de un sujeto en una determinada escala. Así, por ejemplo, se podría preguntar qué eneatipo correspondería a un sujeto que en la prueba del estudio hubiera obtenido 40 puntos. En este caso podríamos ver la tabla anterior de eneatipos y las puntuaciones directas correspondientes, comprobando que el sujeto se encontraría en el eneatipo 8.°, limitado por las puntuaciones directas 36,875 y 42,245. Si no dispusiéramos de esta escala, podríamos transformar la puntuación directa en típica, como sigue:

$X_i = 40$
$\bar{X} = 23{,}45$
$S = 10{,}74$

$$Z_i = \frac{40 - 23{,}45}{10{,}74} = 1{,}54$$

El sujeto se encontraría en el eneatipo limitado por las Z = 1,25 y Z. = 1,75, es decir, en el eneatipo 8.°.

También podríamos aplicar la fórmula de la transformación de Za la escala de eneatipos, sabiendo que en este caso:

$$\text{Eneatipo} = 5 + 2Z = 5 + 2 \cdot 1{,}54 = 8{,}04 \approx 8$$

Es decir, se encontraría en el eneatipo 8.

Como hemos podido observar, por los tres procedimientos llegamos a la misma conclusión. Este ejemplo es extensible a cualquier otro tipo de escala, teniendo en cuenta las características específicas de la misma.

To.2. Teoría de respuesta al ítem

En este apartado presentamos algunas cuestiones relativas a las características, supuestos y demás conceptos básicos y fundamentales de la teoría del rasgo latente.

10.2.1. Ejercicios resueltos

1. Teoría del rasgo latente:

a)Su objetivo y modelos más usuales.

b)Diferencia entre los modelos TRI y LCA. ¿Cuáles son los más desarrollados?

SOLUCIÓN:

a)La teoría del rasgo latente trata de conseguir el objetivo de toda buena medición: hacer inferencias correctas a partir de las respuestas de los sujetos a una prueba o de sus puntuaciones en un test; inferencias sobre las características y rasgos no observables que la prueba o el test pretenden medir.

En la teoría del rasgo latente se supone que, en situaciones de prueba, las respuestas de los sujetos a un test están condicionadas por una serie de rasgos y habilidades (subyacentes) de los propios sujetos. Lo que hace la teoría es establecer un modelo de la relación entre el nivel que la persona posee en el rasgo que se pretende medir y las respuestas que da a los ítems del test.

b)La mayoría de los desarrollos de la teoría del rasgo latente se refieren a los modelos que suponen una distribución "continua" del rasgo o habilidad (0) en la población de sujetos estudiada. Se apoyan en lo que se denomina teoría de respuesta al ítem (TRI).

Otros modelos, LCA (latent class analysis), suponen que la distribución del rasgo o habilidad (0) es discreta.

2. ¿Cuáles son y en qué consisten los supuestos básicos de TRI?

SOLUCIÓN:

Como todo modelo matemático, la TRI, además de especificar las relaciones entre los elementos observables y no observables descritos en el propio modelo, incluye una serie de supuestos sobre los datos a que puede aplicarse el mismo. Destacamos los más importantes:

1.Todos los modelos de TRI suponen que es posible describir matemáticamente la relación entre el nivel que posee el sujeto en un determinado rasgo y sus respuestas a los ítems del test.

2.La mayoría de los modelos funciona bajo el supuesto de unidimensionalidad, lo que significa que hay un solo rasgo o habilidad subyacente en la actuación del sujeto al contestar al conjunto de ítems que forman el test. Decimos la mayoría porque hay también algunos desarrollos multidimensionales.

3.Las respuestas de un sujeto a diferentes ítems en un test son estadísticamente

independientes: este supuesto se conoce con el nombre de "independencia local'.

3. ¿Qué indica la función de respuesta al ítem?

SOLUCIÓN:

La función de respuesta al ítem (FRI) es una función matemática que especifica la relación entre el fenómeno no observable (el nivel del rasgo en el sujeto), y el observado (las respuestas a los ítems).

Expresa que los sujetos con niveles más altos en los rasgos medidos tienen probabilidades esperadas mayores de contestar el ítem correctamente que los examinados que tienen niveles más bajos en tales rasgos.

Si cambia la forma matemática de la relación, surge un nuevo modelo de respuesta al ítem. En el caso de los modelos unidimensionales con ítems dicotómicos, que son los más utilizados, esta función suele denominarse curva característica del ítem (CCI).

4. ¿Cuáles son los modelos unidimensionales (de ítems dicotómicos o no) más usados?

SOLUCIÓN:

Presentamos la síntesis hecha por Hambleton y Swaminathan (1985), y de ellos los más usados son los de uno, dos y tres parámetros logísticos:

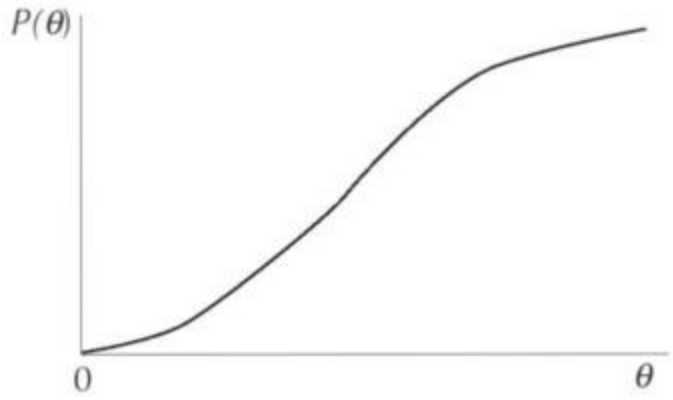
Naturaleza/Datos	*Modelo*	*Referencias*
• Dicotómicos.	• Lineal latente. • Escala perfecta. • Distancia latente. • Uno, dos y tres parámetros. Ojiva normal. • Uno, dos y tres parámetros. Logísticos. • Cuatro parámetros. Logístico.	Lazarsfeld y Henry (1968). Guttman (1944). Lazarfeld y Henry (1968). Lord (1952). Birnbaum (1957, 1958a, 1958b, 1968), Lord y Novick (1968), Lord (1980a), Rasch (1960), Wright y Stone (1979). McDonald (1967), Barton y Lord (1981).
Naturaleza/Datos	*Modelo*	*Referencias*
• Puntuación multicategórica.	• Respuesta norminal. • Respuesta graduada. • Modelo de crédito parcial.	Bock (1972). Samejima (1969). Master (1982).
• Continua.	• Respuesta continua.	Samejima (1972).

5. ¿Qué es la curva característica del ítem (CCI)?

SOLUCIÓN:

Como hemos dicho en el ejercicio 3, la CCI es la FRI para el caso de modelos unidimensionales con ítems dicotómicos.

La curva característica del ítem, por tanto, es la relación funcional entre la probabilidad de respuesta correcta a un ítem y el rasgo latente hipotético. O, dicho en otros términos, la CCI define la probabilidad de respuesta correcta al ítem para un nivel determinado de habilidad en la persona:



-En el eje horizontal se sitúa la escala de habilidad. La más usada tiene de media cero y S = 1 y un recorrido teórico de - ~ a +o. Pero se pueden usar diversas escalas.

-En el eje vertical, la probabilidad de respuesta correcta al ítem, P.(9), donde i es el ítem y j es el nivel de habilidad.

-La curva que relaciona las dos variables es la M.Es la función de regresión no lineal de la puntuación del ítem sobre la habilidad medida por el test.

6. ¿Cuáles son las ecuaciones de los modelos "logísticos" de uno, dos y tres parámetros?

SOLUCIÓN:

$$P_i(\theta_j) = \frac{1}{1+\exp[-(\theta_j - b_i)]} \text{ (de un parámetro)}$$

$$P_i(\theta_j) = \frac{1}{1+\exp[a_i(\theta_j - b_i)]} \text{ (de dos parámetros)}$$

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{1}{1+\exp[-a_i(\theta_j - b_i)]} \text{ (de tres parámetros)}$$

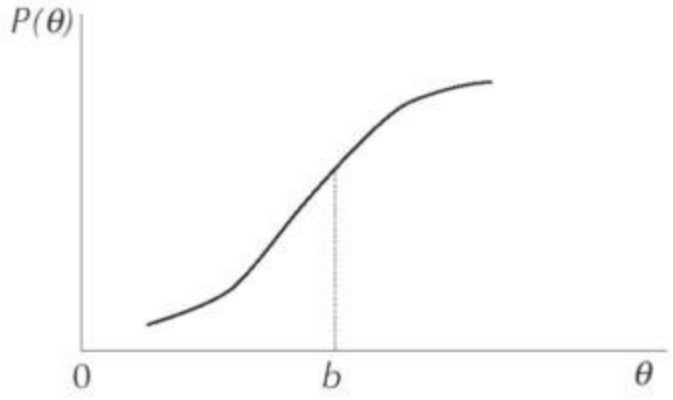
7. El parámetro b:

a)¿Qué significa?

b)¿Qué valores suele tomar?

SOLUCIÓN:

a)El parámetro b expresa la dificultad del ítem. En la CCI, representa el nivel de habilidad correspondiente al punto de inflexión de la curva. Es decir, "localiza" un ítem en la escala de habilidad (por lo que recibe también el nombre de parámetro de "localización").



En un modelo de dos parámetros (c = 0), el valor de b es el punto de la escala de habilidad donde la probabilidad de respuesta correcta es 0,5. Es decir, representa el punto de la escala en que el sujeto tiene un 50% de probabilidades de contestar correctamente al ítem.

b)Cuando las puntuaciones están en una escala de habilidad cuya media es cero y S= 1 (que es la más frecuentemente utilizada), los valores de este parámetro varían

normalmente de -2a+2.

8.En el contexto de la TRI se trabaja, además, con otros dos parámetros de ítem: el a y el c. ¿Qué significado tiene cada uno de ellos?

SOLUCIÓN:

-El parámetro a indica la capacidad de discriminación del ítem. Un ítem que "discrimina bien" entre niveles de habilidad tendrá un valor de a alto.

Este parámetro está relacionado con la "pendiente" de la curva característica en el punto de inflexión. Es el recíproco de la desviación típica de la curva: a = IIS:

•Una curva poco pendiente supone una S numéricamente grande. El ítem discrimina poco, por tanto.

•Una curva bastante pendiente, en cambio, tiene asociada una S pequeña y, por tanto, el ítem discrimina bien entre niveles de habilidad.

Aunque teóricamente el valor de este parámetro puede ir de - 00 a + oo, normalmente los valores están comprendidos entre 0 y 2.

-El parámetro c expresa la probabilidad de que un sujeto con bajo nivel de habilidad (bajo 6) responda correctamente al ítem.

Los modelos de TRI que tienen este parámetro tratan de tener en cuenta, por tanto, la posibilidad de que sujetos con bajo nivel de habilidad tengan ítems correctos con mayor frecuencia de lo esperado bajo el modelo de dos parámetros.

9.Los siguientes son datos correspondientes a los parámetros de varios ítems:

Ítems	*a*	*b*
1	1,2	0,4
2	0,8	1,9
3	1,9	1,4

a)¿Obtendríamos los mismos valores de parámetros si cambiaran las características delgrupo de sujetos en que se obtienen?

b)Si aumentamos el número de ítems a responder, ¿cambiaría el valor de los parámetros?

c)El número de sujetos que haya en cada nivel de habilidad, ¿influye en que el valor de

los parámetros sea diferente?

d)Comparar las respuestas dadas en a), b) y c) con las que daríamos en el contexto de la teoría clásica.

SOLUCIÓN:

Antes de dar las respuestas a estas cuestiones, nos parece oportuno reflejar la base teórica que justificaría todas ellas. Los parámetros de los ítems, en el contexto de la teoría del rasgo latente, tienen una importante característica: son "índices invariantes". No cambia su valor aunque varíen las características de la muestra de sujetos, el número de ítems del test, o ambas cosas.

Según esto, las respuestas serían:

a)Sí, obtendríamos los mismos valores de los parámetros.

b)No cambiaría el valor de los mismos.

c)El número de sujetos que haya en cada nivel no produce cambios en el valor de los parámetros.

d)En la teoría clásica, los parámetros de los ítems "no son invariantes". Su valor depende de la muestra de sujetos, del número de ítems, etc.

10.Observemos los valores de habilidad que hemos asignado a los siguientes sujetos:

Sujetos	θ
1	1,2
2	1,8
3	0,5

a)¿Cambian estos valores si aumentamos el número de sujetos?

b)1 Cambiaría el valor correspondiente al sujeto 2 si los otros dos sujetos tuvieran otros niveles de habilidad?

c)¿Qué ocurre a esos valores de habilidad si los ítems que se usan para obtenerlos cambian?

SOLUCIÓN:

El parámetro del sujeto también es "invariante" (como ocurre con los parámetros de

los ítems). El valor del mismo no depende del número de ítems que responda el sujeto, ni de las respuestas de otros sujetos.

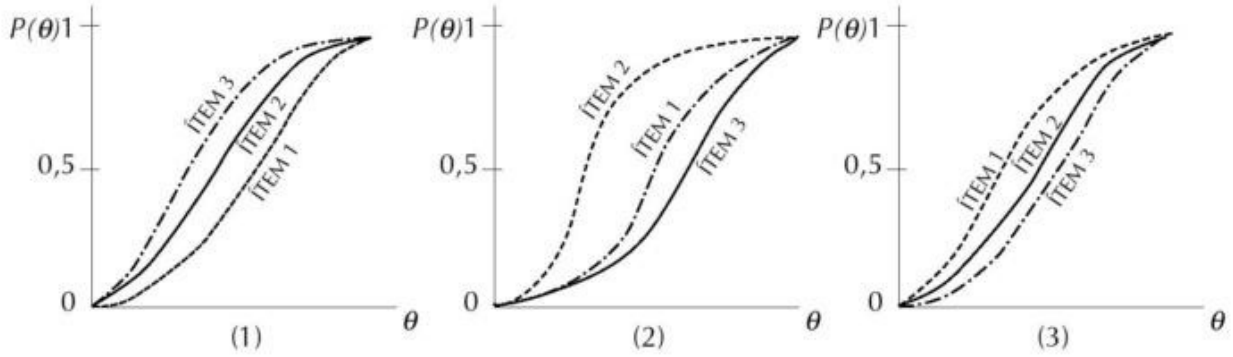
Respondemos, por tanto, a las cuestiones anteriores:

a)No.

b)No.

c)Permanecen igual.

11.Observemos la información que nos proporcionan estos tres grupos de CCL•



a)¿En cuál o cuáles de estos grupos los ítems tienen diferente poder de discriminación?

b)¿Qué ítem tiene mayor poder de discriminación en el grupo (1)? ¿Y en el grupo (2)? ¿Y en el grupo (3)?

c)¿En cuál o cuáles de estos grupos los ítems tienen distinto nivel de dificultad?

d)¿Qué ítem tiene mayor nivel de dificultad en el grupo (1)? ¿Y en el grupo (2)?

e)La probabilidad de respuesta correcta, ¿oscila desde 0 en todos los casos?

f)Si comparamos el ítem 2 de cada uno de los grupos CCI, ¿cuál tiene mayor poder de discriminación? ¿Cuál tiene menor índice de dificultad?

SOLUCIÓN:

a)Si recordamos que el poder de discriminación del ítem (que expresa el parámetro a) está relacionado con la pendiente de la curva, observamos que en los grupos (1) y (3) todos los ítems tienen igual poder de discriminación (sus curvas son igualmente pendientes).

En cambio, en el grupo (2) los ítems difieren en su poder de discriminación (diferente pendiente).

b)En los grupos (1) y (3) todos los ítems tienen igual poder. En el grupo (2), el ítem 2 es el que tiene mayor poder de discriminación. Tiene mucho poder puesto que es muy pendiente.

c)Para responder a esta cuestión, hemos de recordar el significado del parámetro de dificultad b: "nivel de habilidad correspondiente al punto de inflexión de la curva".

Si "localizamos" este parámetro en las CCI de cada uno de los grupos (proyectamos el punto de inflexión de la curva sobre el eje donde se sitúa la escala de habilidad), observamos que, dentro de cada grupo, los b no se ubican en el mismo punto. Es decir, en todos los grupos los ítems tienen diferente nivel de dificultad.

d)Hecha la localización anterior, podemos responder:

-En el grupo (1) el ítem más difícil es el 1, puesto que su parámetro b se localiza más alejado del origen (valor más elevado, por tanto).

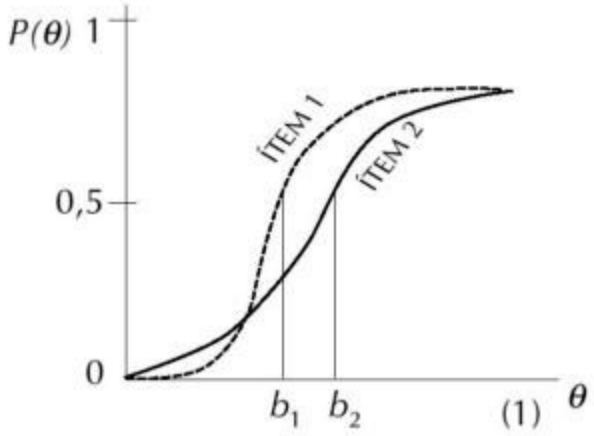
-En el grupo (2) el ítem 3 es el más dificil, por la misma razón.

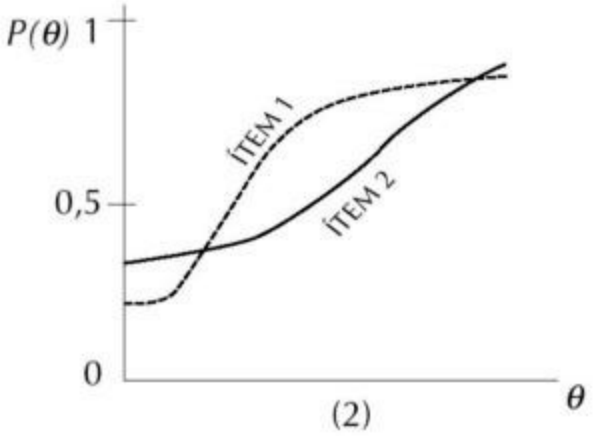
e)Podemos responder que sí, según se observa en el eje correspondiente: P(6).

f)El ítem 2 tiene mayor poder de discriminación en el grupo (2), puesto que es donde tiene mayor pendiente.

En cuanto al índice de dificultad del mismo ítem, observamos que es en el grupo (2) donde tiene menor valor. Es en ese grupo donde dicho ítem se localiza en un punto más bajo (más próximo al origen) de la escala de habilidad (evidentemente, estamos suponiendo, para poder hacer esta comparación, que en los tres grupos de CCI se trata de la misma escala de habilidad).

12.Si nos fijamos en estos dos grupos de CCL-





a)¿Cuál corresponde a ítems de diferente índice de dificultad?

b)¿Cuál corresponde a ítems de diferente discriminación?

c)¿Son modelos lineales?

d)La probabilidad de respuesta correcta, ¿oscila desde 0 en todas las CCI?

e)Si estudiáramos las respuestas de varios ítems, ¿podrían coincidir con las de los ítems aquí representados?

f)¿Dónde se ubicaría el parámetro b de los ítems del grupo (1)?

SOLUCIÓN:

a)Si hacemos la localización del parámetro de dificultad, b, en este caso, según se ha indicado en la respuesta a la tercera pregunta del ejercicio anterior, podemos responder que en los dos grupos de CCI representados aquí, los ítems tienen

diferente índice de dificultad. En el grupo (1) el ítem 2 es el que tiene mayor valor de b; en el (2) tiene mayor índice de dificultad el ítem 1. (En ambos casos, porque el b se localiza en un punto más elevado de la escala de habilidad, según se ha indicado ya en otras ocasiones).

b)También en los dos grupos los ítems tienen diferente poder de discriminación. En el (1), tiene mayor poder el ítem 1 y en el grupo (2), el ítem 2.

c)Evidentemente no, según indican sus gráficas.

d)La respuesta es negativa, según se puede observar en el eje correspondiente: P(6).

e)Sí podrían coincidir. Varios ítems pueden tener el mismo modelo de CCI.

f)Según se indica en la gráfica correspondiente a ese grupo.

13.En el contexto de la TRI, se utiliza un concepto muy importante: la función de información". Las respuestas a las cuestiones que siguen nos ayudan a perfilar el significado del mismo.

a)¿Qué utilidad tiene? ¿De qué nos informa?

b)¿A quién se atribuye la función de información?

c)¿Es la misma para cualquier nivel de habilidad?

d)La contribución de un ítem a la cantidad total de información del test, ¿depende de la contribución de otros ítems?

e)Si comparamos los siguientes datos, ¿qué nos indican?

	$I_{(\theta)}$
$\theta = 0,5$	20,8
$\theta = 2$	11,04

f)¿Qué significa un alto valor de cantidad de información?

g)¿Qué debemos interpretar cuando la cantidad de información es pequeña?

h)¿Hay alguna relación entre la "cantidad de información" que proporciona una serie de ítems (para un nivel de habilidad determinado) y el error con que se estima ese nivel de habilidad?

SOLUCIÓN:

a)La función de información del test o de cada ítem juega un papel clave en la teoría de respuesta al ítem. Nos da a conocer si las medidas que éstos nos proporcionan son eficaces. Nos indica si el test o el ítem de que se trate "estiman bien" el nivel de habilidad que se desea.

b)En la respuesta anterior está implícito algo que destacamos ahora y que responde a la cuestión aquí planteada. Podemos obtener la cantidad de información, para un determinado nivel de habilidad, de cada ítem en particular y del test en conjunto. La cantidad de información de este último es la suma de las que tienen todos sus ítems. En consecuencia, la cantidad de información depende de la calidad y del número de ítems del test.

c)La función de información es diferente y única para cada nivel de habilidad. Cada nivel de habilidad tiene su propia cantidad de información.

d)La contribución que un ítem hace a la cantidad total de información del test es independiente de la contribución de cada uno de los ítems. Es intrínseca a cada ítem, a diferencia de lo que ocurre en la teoría clásica.

e)Dado que la función de información, I(0), es mayor en el caso del nivel de habilidad 0,5 (0 = 0,5), el test o ítem de que se trate "estima mejor" (con más precisión) este nivel de habilidad que el nivel 2.

f)Si un determinado nivel de habilidad (6) tiene alto valor de cantidad de información, significa que esta 6 puede ser estimada por el test o por el ítem con más precisión que otra 0 para la que la cantidad de información sea relativamente baja.

g)Si un nivel de habilidad tiene poca cantidad de información, tenemos que interpretar con precaución ese nivel de habilidad estimada. Sobre todo a la hora de tomar decisiones apoyados en tal información.

h)Evidentemente, es una relación inversa. Basta recordar, para comprender esto, el concepto de función de información aplicado a este caso concreto: la "cantidad de información", que proporciona la serie de ítems aquí considerada, indica si ese conjunto de ítems "estima bien" el nivel de habilidad que se desea.

14. ¿Cuál es el concepto equivalente al defunción de información en TRI, si nos situamos en el contexto de la teoría clásica? ¿Cuáles son algunas de las diferencias que les afectan?

SOLUCIÓN:

El concepto equivalente al de función de información en la teoría clásica es el concepto de fiabilidad. Hay grandes diferencias, sin embargo, entre ambos:

-En la teoría clásica se proporciona un único coeficiente de fiabilidad, que se aplica a todos los sujetos sin tener en cuenta su diferente habilidad. En la TRI se estima el error de medida para cada nivel de habilidad.

-El coeficiente de fiabilidad depende no sólo del test concreto de que se trate, sino también del grupo de sujetos que lo realizan. En cambio, en la TRI la función de información es intrínseca al test; depende sólo de la escala de habilidad y de las CCI de los ítems en el test.

10.3. Teorías de medida: problemas y soluciones

10.3.1. Teoría Clásica. Problemas y soluciones

A) Problemas

1.Un test de 50 elementos se ha aplicado a una muestra representativa de los alumnos de 2.0 de Bachillerato de un distrito madrileño. La correlación entre las dos mitades del test ha sido 0,70. La desviación típica de las puntuaciones es 8.

a)¿Qué coeficiente de fiabilidad tiene el test?

b)¿Qué porcentaje de la varianza de las puntuaciones alcanzadas por los alumnos se debe a la varianza de las puntuaciones verdaderas de los sujetos?

c)¿Qué porcentaje se debe a la varianza de los errores aleatorios?

d)¿Cuál será el intervalo en el que suponemos está la puntuación empírica de 17?

e)¿Qué fiabilidad tendrá el test si le añadimos 10 ítems paralelos?¿ Ysi le eliminamos 12 ítems? Comenta los resultados.

f)¿Cuántos ítems tendríamos que aumentar para conseguir un coeficiente de fiabilidad de 0,90?

2.Hemos construido un test en un grupo cuya variabilidad es 4y conseguimos que el 60% de la varianza total del test se deba a la varianza de las puntuaciones verdaderas. La validez del test es 0,50. La fiabilidad y validez del criterio es 0,70 y 0,55, respectivamente.

a)¿Qué coeficiente de fiabilidad tiene?

b)Si duplicamos su longitud ¿qué coeficiente de validez tendrá? ¿Y si lo reducimos a la mitad?

c)Si conseguimos mejorar la fiabilidad del criterio hasta 0, 80, ¿cuál será la validez del test?

d)¿Cuál sería la validez del test si se consiguieran eliminar los errores de medida tanto del test como del criterio?

ÍTEMS

	1	5	8
ID	0,3	0,2	0,6
IH	0,5	0,4	0,8
IV	0,4	0,5	0,7

e)¿Qué coeficiente de fiabilidad tendríamos que conseguir en el test si quisiéramos que tuviera una validez de 0, 70?

3.Observa estos datos y contesta a las siguientes cuestiones, justificando tus respuestas:

a)¿Cuál de los ítems es más difícil?

b)¿Cuál tiene más consistencia interna con la prueba?

c)¿Cuál contribuye más a medir lo que mide la prueba?

d)¿Cuál mide más lo que deseamos medir?

e)¿Nos hubieran dado necesariamente los mismos datos si los utilizamos en otro grupo de características distintas?

f)Si no tenemos los ID y los IH de todos los ítems, ¿podemos interpretar la consistencia de un ítem con la prueba total?

g)¿Por qué necesitamos la fórmula de corrección para calcular el IH de un ítem?

4.A partir de estos datos:

ÍTEMS

	1	2	3	4	5
ID	0,60	0,80	0,05	0,40	0,20
IH	0,40	0,50	0,10	0,50	0,60
IV	0,30	0,70	0,15	0,20	0,30

a)¿Qué ítem es el más difícil? ¿Por qué?

b)¿Qué ítem explica más varianza del criterio? justifica tu respuesta.

c)¿Qué ítem explica menos varianza del test? justifica tu respuesta.

d)¿Debería eliminarse algún ítem? justifica tu respuesta.

5. A partir de estos datos:

ÍTEMS

	a	*b*	*c*	*d*
ID	0,80	0,10	0,25	0,50
IH	0,60	0,15	0,20	0,40
IV	0,40	0,20	0,30	0,60

a)¿Qué ítem es el más fácil? ¿Por qué?

b)¿Qué ítem explica más varianza del criterio? justifica tu respuesta.

c)¿Qué ítem explica menos varianza del test? justifica tu respuesta.

d)A tu juicio, ¿cuál es el mejor ítem? ¿y el peor? justifica tu respuesta.

6. Un test de razonamiento numérico se ha aplicado a un grupo de sujetos de 3.0 de la ESO aleatoriamente elegidos. A continuación se presentan los resultados obtenidos en el test junto a las calificaciones medias (R) obtenidas en el área de Matemáticas al finalizar los estudios de Bachillerato:

ÍTEMS	SUJETOS										
	1	2	3	4	5	6	7	8	9	10	11
1	+	–	+	–	+	+	+	+	–	+	–
2	+	–	+	–	–	+	–	+	+	+	–
3	+	+	+	–	+	+	–	+	–	+	–
4	–	–	+	+	+	+	+	–	–	+	+
5	+	+	+	–	+	–	–	+	–	–	–
6	+	–	+	+	+	+	–	+	–	+	–
7	–	–	–	+	+	+	–	+	+	+	–
8	–	–	+	+	–	–	–	+	–	+	+
9	+	–	–	–	+	+	–	+	+	–	–
10	+	+	–	+	+	–	+	+	+	+	–
R	20	9	16	12	18	19	5	20	10	19	4

a)¿Qué porcentaje de varianza explica el test del rendimiento en Matemáticas de Bachillerato? Realiza los cálculos pertinentes y justifica tu respuesta.

b)¿Cuántos elementos habría que suprimir del test para que la validez se redujera en 0,1 unidad?

c)¿Qué puntuación directa se le predice a un sujeto de la misma población en Matemáticas, si obtuvo en el test una Z = - 0,2?

d)Suponiendo que el test explicara el 64% de la varianza de otro criterio (B), distinto del que se ha tomado, ¿se podría afirmar que la varianza de error del test ha aumentado? justifica la respuesta y realice los cálculos pertinentes.

e)¿Qué porcentaje de la varianza del test explica el ítem 6? ¿Crees que debería suprimirse? justifica la respuesta.

f)Calcula el centil correspondiente a la puntuación directa X = 3.

7.Un profesor ha elaborado un test de aptitud numérica para alumnos de 4.0 de la ESO. Con el fin de comprobar el valor de dicho test para predecir el rendimiento en Matemáticas de los alumnos en los estudios de este nivel, se tomaron las puntuaciones medias que los alumnos habían obtenido en esta materia a lo largo de los tres cursos (columna B). A continuación se presentan estas puntuaciones así como las que esos mismos alumnos habían o btenido en el test de aptitud numérica.

SUJETOS	ÍTEMS								*B*
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	
1	4	2	3	1	5	1	2	3	20
2	3	1	5	2	4	0	2	1	15
3	4	3	4	0	5	1	3	2	12
4	5	1	3	3	3	1	4	2	18
5	2	0	1	1	2	2	3	1	10
6	3	1	2	1	2	2	3	1	14
7	4	3	3	3	4	2	3	1	18
8	4	2	4	2	3	3	1	1	13
9	5	3	4	3	4	1	5	2	19
10	5	2	3	4	4	1	2	3	18

a)¿Qué proporción de varianza verdadera, respecto de la varianza total, explicaría el test si se incrementase en 6 ítems?

b)Si dos sujetos han obtenido, respectivamente, en el test las puntuaciones 25y 35, ¿existen diferencias significativas entre los resultados de ambos sujetos? (a = 0, 02).

c)¿Se puede afirmar que el ítem d) explica un 40% de la varianza del criterio?

d)¿Cuál será la Z empírica y normalizada que corresponde a la puntuación del sujeto número 3?

e)¿A cuánto asciende la varianza de error del test?

f)Suponiendo que el coeficiente de validez del test fuera 0,70, ¿qué puntuación típica obtuvo un sujeto en el mismo si la puntuación verdadera en el criterio se encuentra entre 18 y

8.Un profesor de 2.0 de Bachillerato construyó una prueba objetiva de su asignatura al objeto no sólo de calificar a sus alumnos, sino también de ver el grado en que su test servía para predecir el éxito de los alumnos en las pruebas de selectividad. Aplicó su test a una pequeña muestra de 10 sujetos, obteniendo los siguientes resultados:

SUJETOS	*a*	*b*	*c*	*d*	*e*	*f*	*H*
1	2	4	1	2	3	5	4
2	3	4	0	3	5	4	3
3	6	4	3	4	6	8	7
4	5	5	4	3	6	9	8
5	5	5	2	2	5	7	6
6	7	5	4	3	5	7,5	8
7	6	4	3,5	3	4	8	7
8	2,5	4	1,5	4	2	3	3,5
9	3	5	1	3	3	4	4,5
10	5	5	4	4	4	6,5	7

Como habrá observado, los ítems no eran de acierto-error, sino pequeños problemas valorables de 0 a 9 puntos cada uno.

Al mes de aplicar la prueba, recogió las calificaciones medias de esta muestra de alumnos en la selectivida d. Tales datos aparecen en la columna H.

a)¿Cuál es el coeficiente de fiabilidad del test de seis ítems construido por el profesor? Utiliza la fórmula de Spearman-Brown.

b)¿Puede decirse que el profesor ha conseguido explicar al menos un 50% de la varianza de las puntuaciones de las pruebas de selectividad?

c)¿Qué puntuación típica obtendrá un sujeto de la población de alumnos de 2.0 de Bachillerato en las pruebas de selectividad, si obtuvo una puntuación directa de 20 puntos en el test elaborado por el profesor? (a = 0, 05).

d)¿Cuál es el índice de validez del ítem b)?

e)¿Puede decirse que el ítem f) explica al menos un 4% de la varianza del test?

9.Un test construido para la selección de personal de un determinado campo profesional se aplicó, para su estudio, a un grupo de sujetos. Disponemos también de las puntuaciones correspondientes a la valoración que de su rendimiento en el trabajo realizó, tras dos años del ejercicio de su actividad (columna R):

SUJETOS	ÍTEMS										*R*
	1	2	3	4	5	6	7	8	9	10	
1	+	+	+	+	+	–	+	+	+	+	20
2	–	+	+	+	–	+	+	–	–	+	14
3	+	+	–	–	+	+	+	+	+	+	18
4	+	–	+	–	–	–	+	–	–	+	11
5	+	+	–	+	+	+	–	+	+	–	16
6	–	–	+	–	+	–	–	+	+	–	10
7	–	–	+	+	+	+	–	–	+	+	14
8	+	–	+	–	+	–	–	+	+	+	12

a)¿En cuántos ítems se deberá incrementar o disminuir el test para conseguir explicar el 64% de la varianza de las puntuaciones en el rendimiento en el trabajo?

b)Si la puntuación verdadera del sujeto en el criterio está comprendida entre una Z1= - 0,75 y una Z. = - 1,25, ¿qué puntuación directa había obtenido el sujeto en el test?

c)¿Qué puntuación Z normalizada le corresponde al sujeto número 5?

d)¿Se puede afirmar que el ítem 4 explica más de un 49% de la varianza del rendimiento en el trabajo?

e)¿Qué proporción de varianza de las puntuaciones del test se debe a la varianza de error? Justifica tu respuesta.

10.A continuación se presentan las puntuaciones obtenidas por un grupo de alumnos de 3.0 de la ESO en una serie de ítems de un test de aptitudes matemáticas y la puntuación media (X) obtenida en este área al final del curso:

SUJETOS	ÍTEMS								X
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	
1	+	–	+	+	+	–	–	+	7
2	+	+	–	+	+	+	–	+	6
3	–	–	–	–	+	–	–	+	3
4	+	–	–	+	–	+	+	–	5
5	+	+	+	–	+	–	–	–	6
6	+	+	–	+	+	+	–	+	8
7	–	–	–	+	+	+	–	+	7
8	+	+	+	+	+	+	+	+	10
9	+	–	–	+	+	+	–	–	3
10	+	+	–	+	–	–	+	–	4

a)¿En qué número de ítems se debería incrementar el test para que explicara un 67% de la varianza del criterio?

b)¿Qué puntuación típica se le predice a un sujeto en el criterio si obtuvo en el test una Xi = 9? (a = 0,02).

c)¿Existen diferencias significativas entre las puntuaciones que el sujeto número 7 obtuvo en el test y en el criterio, sabiendo que la fiabilidad de este último es de 0, 60? (a = 0, 03).

d)Suponiendo que este mismo test se hubiera aplicado a una muestra de 50 sujetos, siendo en este caso la desviación típica igual a 4, ¿cuál sería la fiabilidad correspondiente a dicho test?

e)¿Se puede afirmar que el ítem 6 explica un 40% de la varianza del test? ¿Debería suprimirse este ítem del test? justifica tu respuesta.

f)Calcula la puntuación Z normalizada correspondiente a la puntuación X = 5?

11.Los resultados obtenidos en una prueba de conocimientos que se aplicó a los 10 alumnos al finalizar 3.0 de la ESO se presentan en la siguiente tabla. Conocemos entre otros datos, la correlación entre los ítems pares e impares (r.,= 0,70) y también el promedio de calificaciones de Bachillerato de los mismos alumnos (estas se muestran en la columna B).

SUJETOS	ÍTEMS								*B*
	1	2	3	4	5	6	7	8	
1	2	3	4	3	5	3	4	3	8
2	3	1	2	2	3	0	3	4	6
3	4	5	4	5	3	3	5	4	9
4	1	2	1	0	2	3	3	1	4
5	2	3	3	4	4	4	5	3	8
6	3	2	4	3	3	2	2	1	7
7	5	4	4	4	5	3	3	4	9
8	4	3	2	1	2	1	2	3	6
9	4	4	5	4	3	4	5	3	8
10	1	0	2	2	3	1	2	2	4

a)¿Qué porcentaje de varianza explica la prueba de las calificaciones de Bachillerato? Realiza para ello los cálculos pertinentes.

b)Si se incrementa en 5 el número de ítems, ¿cuál será la nueva validez de la prueba? Y si se eliminaran los errores de medida del test permaneciendo el número original de ítems, ¿cuál sería la validez de la prueba en este caso?

c)¿Quépuntuación típica verdadera se le predice a un alumno de Bachillerato que obtuvo una puntuación directa de 15 en la prueba de conocimientos? (Nivel de confianza del98%).

d)¿Cuál es el penta correspondiente a la puntuación obtenida por el sujeto número 4?

e)A partir de estos datos:

ÍTEMS	*ID*	*IV*	*IH*
1	0,30	0,30	0,40
2	0,60	0,10	0,20
3	0,40	0,60	0,30
4	0,02	0,25	0,10

1.¿Cuál es el ítem más fácil?

2.¿Cuál es el ítem que explica menos varianza del criterio?

3.¿Qué porcentaje de varianza del test explica el ítem número 2?

4.¿Qué ítem o ítems eliminaría de acuerdo con los datos?

NOTA: En todas estas cuestiones justifica la respuesta.

12.En un centro de formación profesional se quiere elaborar una prueba para seleccionar los alumnos que vayan a realizar la especialidad de electrónica.

Los resultados que se presentan en la siguiente tabla recogen las puntuaciones obtenidas por una muestra aleatoria de alumnos en la mencionada prueba. Cada uno de los ítems se valoró de 0 a 5 puntos. La columna C representa las puntuaciones medias que los mismos alumnos obtuvieron al finalizar los estudios de electrónica:

SUJETOS	ÍTEMS								C
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	
1	1	2	1	0	3	2	1	1	9
2	3	4	5	3	3	4	4	5	20
3	4	5	5	5	4	3	4	4	24
4	2	1	2	1	0	3	1	1	10
5	1	2	1	1	0	2	3	1	10
6	1	2	3	1	1	0	2	2	11
7	3	4	3	4	4	5	3	3	22
8	4	5	5	5	4	3	4	4	24
9	5	4	4	4	5	4	3	5	23
10	5	5	4	4	5	4	5	5	29

a)Si se incrementa en 6 elementos el número de ítems, ¿cuál sería la nueva validez del test?

b)¿Cuál es la proporción de varianza de error del test? Interpreta los resultados y saca las conclusiones oportunas.

c)Sise eliminaran los errores de medida del test manteniendo el número inicial de ítems, ¿cuál sería la validez de la prueba en este caso?

d)¿Existen diferencias significativas entre las puntuaciones obtenidas en el test por los sujetos números 3 y 6?

e)¿Existen diferencias significativas entre las puntuaciones que obtuvo el sujeto número 2 en el test y en el criterio?

f) ¿Cuáles son la Z empírica y el eneatipo correspondientes a la puntuación obtenida por el sujeto número 6?

g)¿Se puede afirmar que el ítem número 7 explica un 40% de la varianza del criterio? ¿Consideras que este ítem debería ser suprimido de la prueba? Justifica tus respuestas.

13.Se pretende estudiar el valor predictivo de un test para seleccionar al personal de un grupo de empresas de una misma rama de trabajo. Para ello se aplica el test a un grupo de personas entre las que se selecciona a las más idóneas según los resultados obtenidos. Después de cuatro años de trabajo se efectúa una valoración de la eficacia de cada una de las personas seleccionadas.

A continuación se presentan los resultados en los ítems del test y la valoración posterior (columna Y):

SUJETOS	ÍTEMS										ΣX_i	Y
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*		
1	+	+	+	+	–	+	–	–	+	+	7	12
2	+	+	+	+	+	+	+	–	+	+	9	20
3	–	+	+	+	–	+	+	–	–	–	5	10
4	–	+	+	–	–	+	–	–	–	–	3	6
5	+	+	+	+	+	+	+	+	+	+	10	20
6	+	+	+	+	–	–	+	+	+	+	8	18
7	+	+	–	+	–	+	+	+	+	+	8	16
8	+	+	+	+	–	+	+	+	+	+	9	19
9	–	+	+	–	–	+	+	–	–	+	5	11
10	–	+	+	+	–	+	+	–	–	+	6	12

a)¿Se puede afirmar que el test explica al menos un 80% de la varianza del criterio?

b)Suponiendo que dos sujetos hubieran obtenido en el test unas puntuaciones de Z~ _ - 0,25 y Zi = 0, respectivamente, ¿existen diferencias significativas entre ambas puntuaciones? (a = 0,02)

c)¿Se puede afirmar que el ítem g" explica un 30% de la varianza del test? justifica la respuesta.

d)Calcula la puntuación correspondiente al centil30 y la Z normalizada de la puntuación obtenida por el sujeto 5.

e)Suponiendo que la puntuación predicha en el criterio de un determinado sujeto estuviera comprendida entre Zi = - 0,25 y Z. = 0,25:

1.¿A qué nivel de confianza se realizó la estimación?

2.¿Qué puntuación directa había obtenido el sujeto en el test?

14.En un centro de formación profesional se pretende orientar a los alumnos en función de sus preferencias y de su aptitud para la especialidad correspondiente. Con este fin se elabora, entre otros, un test de aptitudes mecánicas y se les aplica a una muestra de alumnos del centro, seleccionando posteriormente a los más aptos para cursar la especialidad de mecánica. Al finalizar los estudios se compro bó que el test explicaba un 49% de la varianza de los promedios de las puntuaciones obtenidas por los alumnos a lo largo de los cursos. Teniendo en cuenta que los resultados en el test fueron los siguientes:

SUJETOS	ÍTEMS										ΣX_i
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	
1	3	2	0	1	4	2	1	0	1	1	15
2	3	2	1	0	2	5	1	1	3	2	20
3	4	4	2	2	1	5	3	2	2	5	30
4	1	0	1	0	2	1	1	1	0	2	9
5	3	5	3	5	5	5	5	5	4	5	45
6	2	3	2	4	3	3	2	2	2	3	26
7	2	2	1	2	3	2	3	1	1	1	18
8	1	4	3	4	4	4	3	3	2	5	33
9	4	2	4	2	5	5	5	4	4	5	40
10	1	1	1	1	2	2	2	2	1	3	16

Responde a las siguientes cuestiones:

a)Suponiendo que existan diferencias significativas entre las varianzas de las puntuaciones de los ítems pares y de los impares, ¿qué proporción de varianza verdadera es de la varianza total del test?

b)Teniendo en cuenta que este grupo de 10 alumnos es un subconjunto del grupo total de alumnos (N = 80), cuya desviación típica en el test es igual a 8:

1.¿Se puede afirmar que la fiabilidad del test será la misma para todo el grupo que para el grupo de 10 alumnos?

2.Contesta a la misma cuestión anterior, pero refiriéndote a la validez.

c)Si consiguiéramos eliminar los errores de medida del test, ¿cuál sería la proporción de varianza explicada del criterio?

d)¿Qué puntuación se le predecirá a un sujeto en el rendimiento medio de los estudios

de mecánica a un nc del 98%, sabiendo que en el test la puntuación verdadera estimada está comprendida entre Z=-1yZ=I?

e)Calcula las siguientes puntuaciones tipificadas:

1.Centil 30.

2.El eneatipo correspondiente al sujeto 4.

f)Qué porcentaje de varianza del criterio explica el ítem 4?justifica tu respuesta.

15.Se pretende estudiar la validez de un test de aptitud espacial para predecir el rendimiento en los estudios de Arquitectura. Para ello, se selecciona aleatoriamente una muestra de alumnos de 2.0 de Bachillerato que van a ingresar en la Escuela de Arquitectura y se les aplica la mencionada prueba. Al finalizar los estudios, se toma el promedio de las puntuaciones obtenidas por estos alumnos seleccionados a lo largo de todos los cursos (columna H). A continuación se presentan los resultados obtenidos:

SUJETOS	ÍTEMS										*H*
	1	2	3	4	5	6	7	8	9	10	
1	+	+	–	+	–	+	–	+	+	–	8
2	+	+	+	+	–	+	+	+	+	+	9
3	–	+	–	–	–	+	+	–	+	+	2
4	+	–	–	+	+	+	+	+	–	–	6
5	+	+	+	–	+	+	–	+	+	+	7
6	+	+	–	+	+	+	+	+	+	+	8
7	+	+	+	+	+	–	+	+	+	+	8
8	–	–	–	+	–	+	+	–	–	+	4

Suponiendo que la correlación entre ítems pares e impares fuera de 0,65:

a)¿En cuántos ítems se deberá incrementar el test para que la varianza explicada del criterio sea al menos del 70%?

b)Suponiendo que la puntuación verdadera de un sujeto en el criterio estuviera comprendida entre Z. = - 0, 80 y Z. = - 0,10:

1.¿Qué puntuación directa habría obtenido el sujeto en el test?

2.¿A qué nivel de confianza se realizó la estimación?

c)¿Cuál es la proporción de varianza de error del test? justfca detalladamente la

respuesta.

d)Calcula la puntuación Z normalizada y el centil correspondiente a la puntuación obtenida por el sujeto número 4.

e)¿Se puede afirmar que el ítem 5 explica un 49% de la varianza del test?

f)Suponiendo que la fiabilidad del criterio fuera r = 0,70 y que un determinado alumno hubiera obtenido una puntuación directa en el test de 7puntos y en el criterio de 10, ¿existen diferencias significativas entre ambas puntuaciones?

B) Soluciones

1.a) R.,= 0,82.

b)0,82 (82%).

c)(1 - 0,82 = 0,18) x 100 = 18%.

d)Límite superior: 23,58. Límite inferior: 10,42.

e)R x = 0,845 (si se aumenta en 10 ítems).

Rx0,77 (si se reduce el test en 12 ítems).

f)Habría que aumentar el test en aproximadamente 44 ítems.

2.a) 0,60 (coeficiente de fiabilidad).

b)R = 0,56 (si duplicamos su longitud).

R = 0,42 (si reducimos el test a la mitad).

c)Ry=0,53.

= 0,78.

e)Nó se podría lograr esa validez aumentando sólo la fiabilidad.

3.a) El 8 (ID = 0,6, valor más alto).

b)El 8 (IH= 0,8, valor más alto de los tres).

c)El 8 (IH= 0,8, valor más alto de los tres).

d)El 8 (IV = 0,7, valor más alto de los tres).

e)No.

f)No.

g)Porque el número de elementos es pequeño.

4.a) El 3, por su menor índice de dificultad.

b)El 2 (IV= 0,7).

c) E13 (IH= 0,1).

d)Sí, el 3 (IV e IH muy bajos).

5.a) El a (ID= 0,80)

b)El d, porque es el que más correlaciona con el criterio (IV= 0,60).

c)El b, el que tiene menor índice de homogeneidad (IH= 0,15).

d)El mejor: el d

e)El peor: el b.

6.a) 91,98%.

b)Aproximadamente se debería reducir en 4 ítems.

c)Entre 9,16 y 16,22 puntos.

d)No, ya que el coeficiente de fiabilidad no ha variado.

e)Aproximadamente el 50% de varianza (corrección IH) Por este criterio no debería suprimirse, pero hay que tener en cuenta la validez. En este caso, el IVdel ítem es muy alto (0,89), por tanto, no debe suprimirse.

f)C19•

7.a) El coeficiente de fiabilidad alcanzado: R.= 0,757.

b)Existen diferencias significativas (Zmp - 3,07).

c)No, explica menos (36%).

d)Z = 0,383- e

La Z calculada a partir de la fórmula del centil es igual a 0,25 y a partir del porcentaje acumulado - correspondiente al límite superior de la puntuación (22,5) - es igual a 0,52. Esta discrepancia se acusa especialmente por el reducido número de sujetos con que trabajamos en el ejercicio.

e)En términos de proporción: 0,31; en términos de cantidad de varianza: Sé = 5,38.

f)Z=2,37.

8. a)RXX = 0,978.

b)Si, ya que explica más del 90% (0,9672. 100).

c)Entre Z. = 0, 5 y Z. = - 1,07.

d)IV= 0,50.

e)Sí, explica más: 0,452 (calculado con la corrección del índice de homogeneidad).

9.a) Habría que eliminar aproximadamente cuatro ítems.

b)X. = 6,65.

c)Por la fórmula del centil, Z = 0,49 y por porcentaje acumulado - correspondiente al límite superior de la puntuación 7 (LS = 7,5) - Z = 0,67. Esta discrepancia se acusa especialmente por el reducido número de sujetos de este ejercicio.

d)No, explica aproximadamente un 25% de la varianza R.

e)P = 0,73.

10.a) En aproximadamente un ítem.

b)Límite inferior: 0,74; límite superior: 3,58.

c)No existen diferencias significativas (Z.,P = 0,98). e

d)zx 0,89.

11e) No, explica menos, aproximadamente un 4%: (0,206. 100) (con corrección de IM. Por este criterio no sería necesario suprimirlo, aunque habría que tener en cuenta su índice de validez. En este caso, el IV es de 0,308, por lo que no debería suprimirse.

f)Por la fórmula del centil, Z = 0,25 y por porcentaje acumulado - correspondiente al límite superior de la puntuación 5 (LS = 5,5) - Z = 0,52. Esta discrepancia se acusa especialmente por el reducido número de sujetos de este ejercicio.

11.a) 92,16%: r=0,962)y

b)Ry = 0,98.

c)Entre Z~ 1,535 y Z. = - 3,685.

d)Penta 2.

e)1. El ítem 2, ya que su índice de dificultad tiene el valor más alto (ID = 0,60), lo que indica que ha sido resuelto correctamente por bastantes sujetos.

2.El ítem 2 (IV= 0,10).

3.4% de la varianza del test (0,22 = 0,04).

4.El ítem 2, por su bajo índice de validez y quizás el ítem 4 por su bajo índice de homogeneidad, si bien el índice de validez es aceptable.

12. a) Ry=0,987.

b)1 y 0,984 = 0,016.

c)R=0,99.

d)Existen diferencias significativas (a= 0,01) (Znp = 11,92).

e)No se puede saber, ya que se desconoce la fiabilidad del criterio.

f)Z = - 1,13.

Eneatipo 3.

g)Explica aproximadamente el 53% de la varianza del criterio (0,732).

13.a) Sí, ya que explica el 94,09% (X~y=0,972).

b)No existen diferencias significativas al a = 0,02 (Zmp = 0,47).

c)No, ya que explica el 8,4% (IH= 0,29, con corrección).

d)C30 = 5,5.

Por la fórmula del centil, Z = 1,65; por porcentaje acumulado - correspondiente al límite superior de la puntuación 10 (LS = 10,5) - no se puede obtener, ya que la distribución normal es asintótica con el eje de abscisas y no hay ningún valor que deje por debajo de sí el 100% de los casos. Esta discrepancia se debe al reducido número de sujetos.

e)1. NC= 70,16%.

2.X= 7.

14.a) 0,8: proporción de la varianza verdadera que es de la varianza total (calculado por el procedimiento de Guttman).

b) 1.No,rr=0,617.

.

d)Entre=y=1,65.Z-1,65Z

e)C30=16,83.

Eneatipo 2.

f)Con los datos que poseemos no se puede saber, ya que faltan las puntuaciones de cada sujeto en el criterio.

15.a) Se deberá incrementar en aproximadamente tres ítems.

b) 1. X=6,01.

2. NC= 44,48%.

c) Se = 1 - 0,78 = 0,22.

d)Por la fórmula del centil =-0,32; por porcentaje acumulado - correspondiente al límite superior de la puntuación 6 (LS= 6,5) - es igual a 0 (discrepancia acusada por el reducido número de sujetos del ejercicio).

e)No explica el 49% de la varianza del test (IH= 0,23, aplicada la corrección de dicho índice).

f)Sí, existen diferencias significativas al a = 0,05 y no al a = 0,01 (Z.,P.

10.3.2. Teoría de respuesta al ítem: problemas y soluciones

A) PROBLEMAS

1.¿Cómo se llaman los parámetros de ítem que se estiman en un modelo logístico de tres parámetros?

2.En el contexto de la teoría del rasgo latente, indica lo que significan los siguientes símbolos: (e;),1(8),á;.

3.Si en un test, el contestar a un ítem condiciona la respuesta a otro, ¿se da el supuesto de independencia local?

4.Indica algunas de las aplicaciones que los modelos y desarrollos de la teoría del rasgo latente tienen en el contexto de la medida.

5.Las estimaciones de los parámetros del ítem y ¿son también invariantes como lo son los valores verdaderos de los parámetros?

6.Indica la diferencia entre función de respuesta del ítem (FRI) y curva característica del ítem (CCI).

7.La estimación del rasgo o habilidad 0, ¿es también invariante como lo es el valor verdadero del parámetro (0)?

8.En términos de ecuación de regresión, ¿qué representa la curva característica del ítem?

9.Observa las siguientes respuestas de una muestra de sujetos a una serie de ítems y contesta a estas cuestiones:

a)¿Cuál es el patrón de respuesta del sujeto 2?

b)¿Cómo se ha contestado el ítem b por los diferentes sujetos?

ÍTEMS	SUJETOS					
	1	2	3	4	5	6
a	1	1	0	1	0	1
b	0	1	0	0	1	0
c	1	0	1	1	1	0
d	1	1	1	0	0	0
e	0	1	0	1	1	1
f	1	0	1	1	0	1

10.¿Cuáles son los valores posibles y los más frecuentes de la escala de habilidad?

11.Estos datos corresponden a los parámetros de varios ítems:

ÍTEMS	*a*	*b*
1	0,9	1,0
2	1,5	0,5
3	0,3	1,5
4	1,2	–2,0

a) ¿Cuál es el ítem más dificil?

b) ¿Cuál es el más fácil?

C) ¿Cuál es el que tiene mayor poder de discriminación?

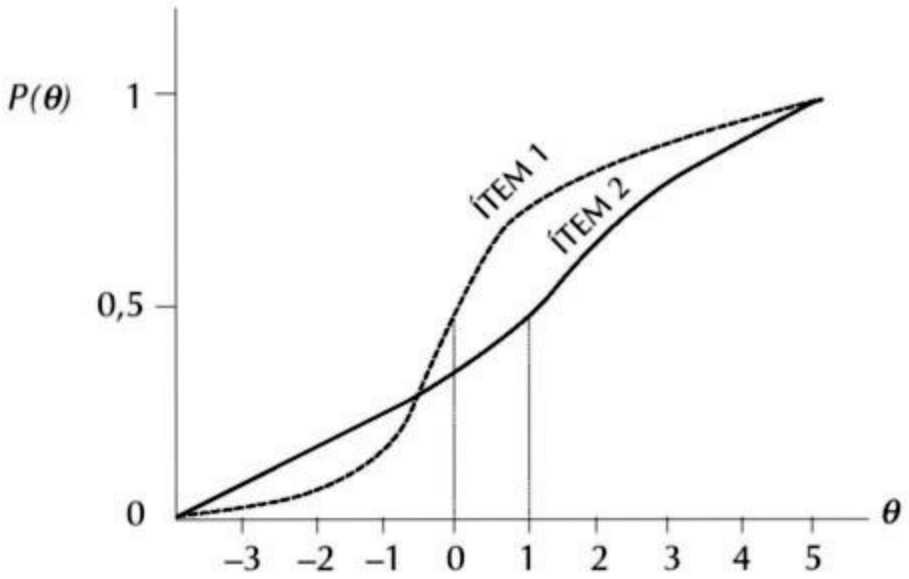
d)Si representáramos las CCI correspondientes a estos ítems, ¿cuál sería la más plana? ¿cuál la más pendiente?

12.Uno de los conceptos utilizados en el contexto de TRI es el de función característica del test":

a)¿Qué significado tiene?

b)¿Para qué se utiliza?

13.Si nos fijamos en estas CCI:



a)¿Qué ítem es más fácil?

b)¿Qué ítem tiene mayor parámetro de discriminación?

c)¿Dónde se ubicaría el parámetro b de cada ítem?

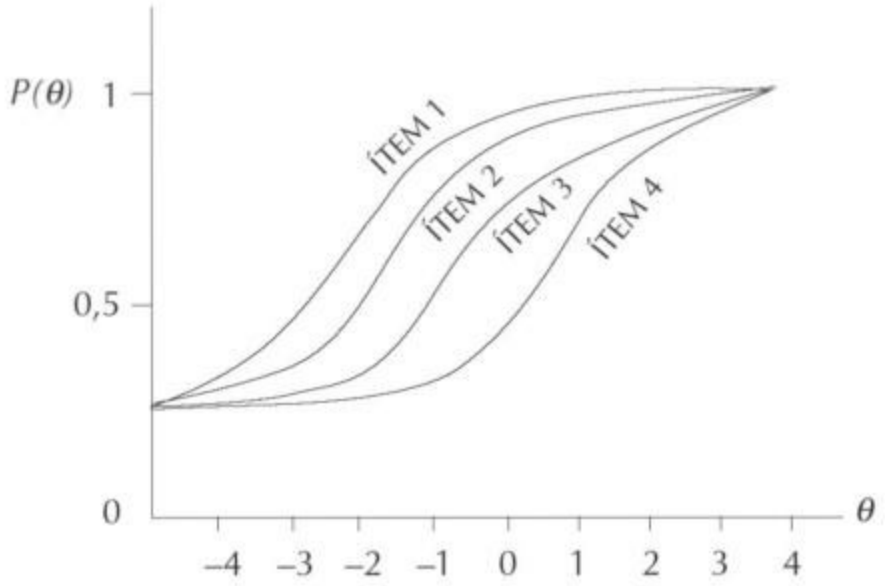
d)¿Cuál es el valor (aproximado) del parámetro de dificultad de los dos ítems?

14.Supongamos una CCI cuya forma general sea P(O) = b. + ae. (lineal como se observa).

a)Si el parámetro de discriminación del ítem esa = 0,80 y el nivel de habilidad de los sujetos es 0 = 2,5, ¿cuál es la probabilidad de respuesta correcta en el caso de que el parámetro de dificultad sea b = - 0,35 y - 0,15, respectivamente. Comenta los resultados.

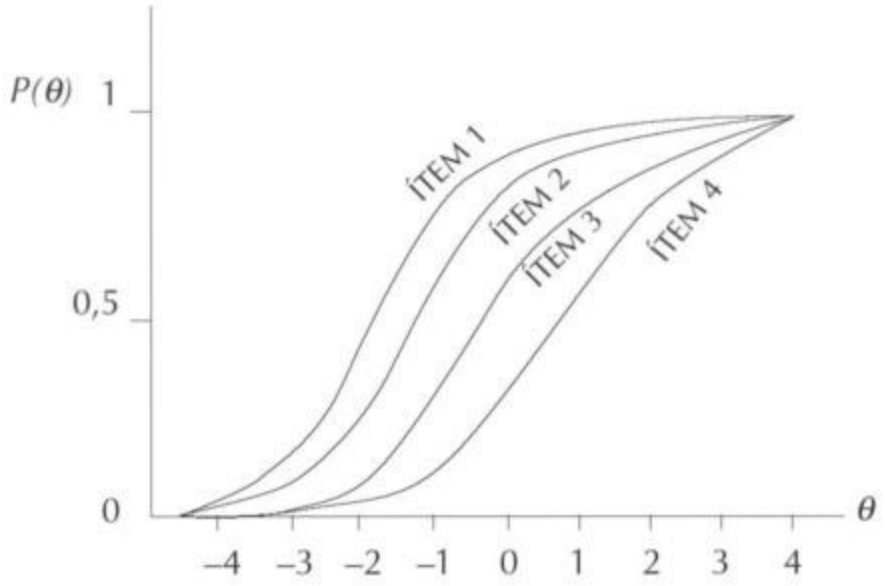
b)Si el parámetro de dificultad del ítem es -b = - 0,15, calcula la probabilidad de respuesta correcta para los niveles de habilidad 9 = 1 y 9 = 2,5 en el caso de que sea a = 0,15. Comenta los resultados.

15.A continuación aparece una representación gráfica de cuatro M.Los parámetros correspondientes son: b = -2, - 1, 0, 1; a =1,39y e = 0,25.



a)Con la simple observación de esta gráfica, ¿podemos saber la probabilidad asociada de respuesta correcta a los niveles de habilidad correspondientes a los parámetros de dificultad de cada uno de los ítems?

b)¿Y en esta otra representación gráfica en la que los parámetros son b = -2, - 1, 0, 1; a = 1,39; c = 0?



c)¿Se podría decir que las CCI del primer ejemplo y las del segundo están asociadas al mismo tipo de modelo? Justifica la respuesta.

B) Soluciones

1.El parámetro de discriminación (a.), el de dificultad (b) y el de adivinación o asíntota baja (c).

2.P. (O): Probabilidad de que un sujeto con nivel de habilidad j conteste al ítem i correctamente.

I (0): función de información del test o de un ítem.

á.: estimación del parámetro de discriminación.

3.No.

4.Construcción y rediseño de tests, selección de ítems sesgados, tests a la medida, etc.

5.No son invariantes, porque se han estimado a partir de muestras de sujetos y éstas fluctúan.

6.La función de respuesta del ítem especifica la relación entre el nivel de habilidad del sujeto (rasgo subyacente) y la respuesta al ítem. Curva característica del ítem es como se suele llamar a la función de respuesta al ítem cuando estamos en el contexto de modelos unidimensionales con ítems dicotómicos. Ambas se representan por P. (e).

7.No es invariante porque se estima a partir de una muestra de ítems y éstos fluctúan.

8.Representa la regresión de las puntuaciones del ítem sobre la habilidad de los sujetos.

9.a) 1 1 0 1 1 0

b) 1 0 0 1 0

10.Los posibles están entre--y+

Losnormales desde - 3 a + 3.

11.a) E13.

b)E14.

c)E12.

d)La más plana es la del 3. La más pendiente la del 2.

Justificación: a es inversamente proporcional a la variabilidad.

12.a) Es el promedio de las funciones de respuesta al ítem para n ítems.

b)Para las transformaciones de la escala de habilidad.

13.a) El 1 (como se observa si localizamos el b).

b)El 1 (más pendiente, menos variable).

c)Según se indica en la gráfica.

d)b1 0; b2 = 1.

14.a) P(O) = 1,65 y 1,85, respectivamente. Observamos que a medida que aumenta el valor de b, aumenta la probabilidad de respuesta correcta.

b)

	$a = 0{,}35$	$a = 0{,}15$
$\theta = 1$	$P(\theta) = 0{,}2$	$P(\theta) = 0$
$\theta = 2{,}5$	$P(\theta) = 0{,}725$	$P(\theta) = 0{,}225$

La diferencia (entre la probabilidad de respuesta correcta de dos niveles de habilidad) aumenta al aumentar el valor de a: 0,725-0,2 > 0,225-0. Este es un dato empírico de que el poder de discriminación del ítem aumenta cuando a es mayor.

15.a) Probabilidad mayor que 0,5, puesto que es un modelo que tiene parámetro c.

b)Probabilidad igual a 0,5, puesto que se trata de un modelo de dos parámetros.

c)No. Las del primero corresponden a uno de tres parámetros y las del segundo no.

Bibliografía

ARRUGA 1 VALERI, A. (1987): Introducción al test sociométrico, 4.a ed., Herder, Barcelona.

BEGGS, D. L. y LEWIS, E. L. (1979): Evaluación del proceso educativo, TEA, Madrid.

BELL, H. M. (1978): Cuestionario de adaptación para adolescentes, Adaptación española de Cerdá, E., Herder, Barcelona.

BIGGS, JOHN (2008): Calidad del aprendizaje universitario, Narcea, Madrid.

BLOOM, B. S. (1971): Taxonomía de los objetivos de la educación. Ámbito cognoscitivo, Marfil, Alcoy.

BLANCO, A (COORD.) (2009): Desarrollo y evaluación de competencias en Educación Superior, Narcea, Madrid.

BROWN, S. y GLASNER, A. (Ed.) (2007): Evaluar en la universidad, Narcea, Madrid.

BRUECKNER, L. S. y BOND, G. L. (1971): Diagnóstico y tratamiento de las dificultades del aprendizaje, Rialp, Madrid.

CAMPBELL, D.T. y STANLEY, J. C. (1978). Diseños experimentales y cuasiexperimentales en la investigación social, Amorrortu, Buenos Aires.

CARREÑO, F. (1977): Instrumentos de medición del rendimiento escolar, Trillas, México.

COOK, T. D. y CAMPBELL, D.T. (1979). Cuasi-experimentation, Rand McNally, Chicago.

COOK, T. D. y REICHARDT, CH. (1986). Métodos cualitativos y cuantitativos en investigación, Morata, Madrid.

CRONBACH, L. J. (1971): Fundamentos de exploración psicológica, Biblioteca Nueva, Madrid.

FERNÁNDEZ CANO, A. (2005). Métodos para evaluar la investigación en psicopedagogía, Síntesis, Madrid.

FERNÁNDEZ DÍAZ, M.J.; GARCÍA RAMOS, J.M.; ASENSIO, I. y CARBALLO, R. (2011): 225 Problemas de estadística aplicada a la Educación, Síntesis, Madrid.

FERNÁNDEZ DÍAZ, M.J.; GARCÍA RAMOS, J.M.; FUENTES, A. y ASENSIO, I. (1990): Resolución de problemas de estadística aplicada a las ciencias sociales, Síntesis, Madrid.

(1990): 225 Problemas de estadística aplicada a las ciencias sociales, Síntesis, Madrid.

GAGNE, R. M. (1972): Planificación de la enseñanza, Trillas, México.

GARCÍA HOZ, V. (1976): Manual de test para la escuela, Escuela Española, Madrid.

GARCÍA HOZ, V. y PÉREZ JUSTE, R. (1984): La investigación del profesor en el aula, Escuela Española, Madrid.

GARCÍA RAMOS, J. M. (1984): Validación y medida del constructo dependencia-independencia de campo perspectivo, Tesis doctoral inédita, UCM, Madrid.

(1986): "Validación de constructo en el ámbito pedagógico", Revista Española de Pedagogía, 174, págs. 535-554.

(1987): "Calificaciones escolares", en Diccionario de Didáctica y Tecnología de la Educación, Anaya, Madrid, págs. 37-41.

(1989 a): Bases pedagógicas de la evaluación, Síntesis, Madrid.

(1989 b): "Evaluación de Programas, Evaluación de Centros Educativos", en Pérez Juste, R. y García Ramos, J. M. (1989), Diagnóstico, evaluación y toma de decisiones, Rialp, Madrid.

(1991): Recursos metodológicos en la evaluación de programas, Bordón, Vol. 43, 4 (pp. 461-476).

(1999):"Investigación y evaluación. Implicaciones y efectos", Revista Complutense de Educación, Vol. 10, n.° 2: 189-214.

GARCÍA YAGÜE, J. y otros (1975): Todos los tests empleados en España, TEA-MEPSA, Madrid.

GARRISON, J. (1986). Some principles ofpostpositivistic philosophy of science, Educational Researcher (pp. 9-15).

GRONLUND, N. (1973): Medición y evaluación en la enseñanza, Trillas, México.

GUBA, E. G. y LINCOLN, 1. S. (1989): Fourthgeneration evaluation, Sage, Newbury Park.

HOWE, K. y EISENHART, M. (1993): "Criterios de investigación cualitativa (y cuantitativa), Prolegómenos", Revista de Educación, 33 (pp. 173-189).

HUSEN, T (1988): "Paradigmas de la investigación en educación. Un informe del estado de la cuestión", en Dendaluce, 1. (Ed.). Aspectos metodológicos de la investigación educativa (pp. 46-60), Narcea, Madrid.

KERLINGER, E N. (2005) 5.a ed.: Investigación del comportamiento. Técnicas y metodología, Interamericana, México.

KNIGHT, P. T. (2008): El profesorado de educación superior, Narcea. Madrid.

KRATWORLD, R. (1971): Taxonomía de los objetivos de la educación, Marfil, Alcoy.

LANDSHEERE, V. y G.DE (1977): Objetivos de la educación, Marfil, Alcoy.

LÁZARO, A. y ASENSI, J. (1979): Vademecum de pruebas psicopedágogicas, MEC, Madrid.

LÁZARO, A. y ASENSI, J. (1987): Manual de Orientación y Tutoría, Narcea, Madrid.

LÓPEZ PASTOR, V. M. (Coord.) (2009): Evaluación formativa y compartida de Educación Superior, Narcea. Madrid

LÓPEZ NOGUERO, E (2007): Metodología participativa en la enseñanza universitaria, Narcea. Madrid.

LORD, F. M. y NOVICK (Eds.) (1968): Statistical theories of mental test scores, Reading, Mass.: Addison-Wesley.

MAGER, R. E (1982): Formulación operativa de objetivos didácticos, 6.a ed., Marova, Madrid.

MARTÍNEZ ARIAS, M.R.; HERNÁNDEZ LLOVEDA, M.J.; HERNÁNDEZ LLOVEDA, M. C. (2006): Psicometría, Alianza Editorial, Madrid.

MORALES VALLEJO, P. (1988): Medición de las actitudes en Psicología y Educación, Ttarttalo, San Sebastián.

MUÑIZ, J. (1992): Teoría clásica de los tests, Pirámide, Madrid.

(1993): Teoría de respuesta al ítem, Pirámide, Madrid.

NUNNNALLY, J. C. (1978). Psychometric theory, McGraw-Hill, New York.

ORDEN, A.DE LA (1969): "La evaluación del rendimiento educativo y la calidad de la enseñanza", Revista de Educación. N.° 206.

(1981): "Evaluación del aprendizaje y calidad de la educación." En Varios, La Calidad de la educación, CSIC. Madrid,

(1982): La evaluación educativa, Docencia, Buenos Aires.

(1984): "Investigación pedagógico-experimental y praxis educativa", en VV.AA., Cincuentenario de los estudios de Pedagogía. Madrid: Facultad de Filosofía y CC. de la Educación. Universidad Complutense de Madrid.

(1985): Hacia una conceptualización del producto educativo, Ponencia en III Seminario de Modelos de Investigación en Educación, Gijón.

(1985): (Ed.) "Investigación Educativa", Diccionario de CC. de la Educación. Anaya, Madrid.

(1989): Investigación cuantitativa y medida en educación, Bordón, vol. 41, 2 (pp. 217-236).

(1995):Innovación e Investigación en el ámbito educativo, Bordón, 47 (2) (pp. 135-141).

(2011): El problema de las competencias en la educación general, Bordón. Vol 63, n.° 1 (pag. 47-76).

ORDEN, A.DE LA, ASENSIO, L; CARBALLO, R.; FERNÁNDEZ, M.J.; FUENTES, A.; GARCÍA RAMOS, J.M.; GUARDIA, S. y NAVARRO, M. (1997): "Desarrollo y validación de un modelo de calidad universitaria como base para su evaluación", Revista Electrónica de Investigación y Evaluación Educativa. Vol. 3. N.° 2.

PÉREZ JUSTE, R. (1976). Elaboración y validación de una prueba de adaptación social, Tesis Doctoral Inédita, UCM, Madrid.

(1983): La medida en educación, Unidades didácticas, UNED, Madrid.

(1986): Evaluación de los logros escolares, UNED, Madrid.

PHI DELTA KAPPA (1971), en STUFFLEBEAN, D.: Educational evaluation and decisión making, Itaca Peacock, Illinois, págs. 40-47.

POLAINO, A. (2010): Antropología e investigación en las CC Humanas, Instituto de Investigaciones Económicas y Sociales. Universidad Francisco de Vitoria, Unión Editorial, Madrid.

POPHAM, W. J. (1980): Problemas y técnicas de la evaluación educativa, Anaya, Madrid.

(1983): Evaluación basada en criterios, Magisterio Español, Madrid.

RODRÍGUEZ SANMARTÍN, A. (1979): Fundamentos y estructura de la evaluación educativa, Anaya, Madrid.

ROTGER AMENGUAL, B. (1984): El proceso programador en la escuela, Escuela Española, Madrid.

SHADISH, W. R. (1989): "Critical mutiplism: a research strategy and its attendant tactics", en Sechrest, L. et al. (Eds.), Health services research: a fous on AIDS. Agency health care policy res. Rockville, M.D. (pp. 5-28).

SHULMAN, L. (1988): "Paradigms and research programs in the study of teaching. A contemporary perspective", en Wittrock, M. C. (Ed.), Handbook of research on teaching, McMillan, New York. Traducido al castellano. Paidós/MEC, 1989 (pp. 9-84).

STUFFLEBEAN, D. L. y SHINKFIELD, A. J. (1987): Evaluación sistemática. Guía teórica y práctica, Paidós/MEC, Madrid.

SUMMERS, G. F. (1976): Medición de actitudes, Trillas, México.

TENBRICK, T. D. (1981): Evaluación. Guía práctica para profesores, Narcea, Madrid.

THORNDIKE, R. L. y HAGEN, E. (1975): Tests y técnicas de medición en psicología y educación, Trillas, México.

WEIS, C. H. (1975): Investigación evaluativa, Trillas, México.

WITTROCK, M. C. (1986): Handbook or research on teaching, McMillan, New York, (3.a ed.) (Traducción en Paidós/MEC).

WOLF, R. M. (ed.) (1987): "Educational evaluation: the state of the field", International journal of Educational Research (número monográfico sobre evaluación), vol. 11 (1).

(1990):Evaluation in education, PraegerPb., 3.a ed. New York.

Índice

Introducción 13
1. Investigación y evaluación. Implicaciones y efectos 18
1.2. Ciencia 20
1.4. Dos conceptos: causalidad y empirismo 22
1.5. El método y la cuestión de los paradigmas en la investigación y evaluación educativas 23
1.6. La medida en investigación y evaluación educativas 27
1.7. La evaluación educativa y sus ámbitos 28
1.8. Investigación e innovación educativa 29
1.10. Investigación sobre evaluación. Algunas conclusiones 35
2. ¿Qué es la evaluación? Necesidad de evaluar 41
2.1. La evaluación y su evolución conceptual 50
2.2. Evolución del concepto de evaluación 57
2.3. Algunas conclusiones sobre la evaluación y su incidencia en la calidad educativa 58
3. El rendimiento educativo. Acotaciones 60
3.2. Evaluación del rendimiento: formas e instrumentos 65
4. Modalidades de la evaluación 68
4.1.2. Evaluación continua 80
4.1.3. Evaluación final 83
4.2. Evaluaciones formativa y sumativa. Funciones de la evaluación 86
4.3. Evaluaciones normativa, criterial y personalizante (sistemas de referencia en la evaluación) 88
4.4. Otras modalidades de evaluación 93
4.5. Otras funciones de la evaluación 94
5. La medida en la evaluación educativa 96
6. Instrumentos de registro de datos 112
6.1.1. Observación directa 114

6.1.2. Observación indirecta 121
6.2. Instrumentos de encuesta 124
6.2.1. El cuestionario 125
6.2.2. La entrevista 127
7. Instrumentos de medición: tests y pruebas 129
7.1. Instrumentos externos de medida (tipificados) 133
7.1.2. Tests de personalidad, intereses, adaptación, actitudes, estilos cognitivos 135
7.1.3. Tests e instrumentos externos de medición del rendimiento y de otras áreas pedagógicas 146
7.1.4. Valoración e interpretación de los tests tipificados 149
7.2. Instrumentos internos de medida del rendimiento 156
7.2.1. Referencia criterial y normativa 157
7.2.2. La prueba objetiva 160
7.2.3. Problemas 179
7.2.4. La prueba de ensayo 181
8. La práctica evaluadora en el centro escolar 188
8.2. Pasos de un proceso tipo de evaluación 192
8.2.1. Fase de planificación 194
8.2.2. Fase de construcción y selección de instrumentos de recogida de información 199
8.2.4. Evaluación (formulación de juicios, toma de decisiones e informes) 202
9. Teorías de la medida 212
9.1.1. La Teoría de la puntuación verdadera 214
9.1.2. Algunas limitaciones de la Teoría clásica de la medida 216
9.2. Teoría de respuesta al ítem 217
10. Resolución de problemas sobre medición en educación 225
10.1. Teoría Clásica de la medida 228